Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной международной конференции «Диалог» (2016)

Выпуск 15

Computational Linguistics and Intellectual Technologies

Proceedings of the Annual International Conference “Dialogue” (2016)

Issue 15

УДК 80/81; 004
ББК 81.1
К63

Программный комитет конференции выражает
искреннюю благодарность Российскому фонду
фундаментальных исследований за финансовую поддержку

Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 1–4 июля 2016 г.). Вып. 15 (22). — М.: Изд-во РГГУ, 2016.

Сборник включает 68 докладов международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2016», представляющих широкий спектр теоретических и прикладных исследований в области описания естественного языка, моделирования языковых процессов, создания практически применимых компьютерных лингвистических технологий.

Для специалистов в области теоретической и прикладной лингвистики и интеллектуальных технологий.

Предисловие

15-й выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит избранные материалы 22-й международной конференции «Диалог». На основании мнений нашего рецензентского корпуса для публикации в ежегоднике редсоветом было отобрано 68 докладов из числа примерно ста работ, которые были рекомендованы по результатам рецензирования для представления на конференции в 2016 году.

Работы в сборнике отражают те направления исследований в области компьютерного моделирования и анализа естественного языка, которые по традиции представляются на конференции:

- Компьютерные лингвистические ресурсы
- Компьютерный анализ документов (классификация, поиск, анализ тональности и т. д.)
- Корпусная лингвистика (создание, разметка, методики применения и оценка корпусов)
- Лингвистические онтологии и автоматическое извлечение знаний
- Лингвистический анализ Social media
- Лингвистический анализ речи
- Машинный перевод текста и речи
- Модели и методы семантического анализа текста
- Модели общения
- Теоретическая и компьютерная лексикография
- Типология и компьютерная лингвистика
- Формальные модели языка и их применение в компьютерной лингвистике

В соответствии с традициями «Диалога», старейшей и крупнейшей конференции по компьютерной лингвистике в России, отбор работ основывается на представлении о необходимости соединения новых методов и технологий анализа языковых данных с серьезными лингвистическими исследованиями. Одной из важнейших целей конференции была и остается поддержка создания современных компьютерных ресурсов, моделей и технологий для русского языка.

В рамках годового цикла проведения конференции реализуется программа специального направления Dialogue Evaluation — тестирований технологий решения отдельных задач компьютерного анализа русского языка. На «Диалоге» по традиции подводятся итоги проведенных тестирований, а статьи организаторов и наиболее успешных участников представляются в настоящем сборнике.

В этом году было проведено три таких мероприятия:

- продолжился цикл тестирований различных подходов к анализу тональности текстов;
- впервые для русского языка было проведено сравнение систем автоматического исправления опечаток;
- тестировались методы автоматического извлечения сущностей и фактов.

Значимость дорожек Dialogue Evaluation очень велика: в результате их проведения не только выявляется реальный уровень технологий участников, но и создается основа для сравнительной оценки эффективности результатов для всего направления, прежде всего в виде т. н. золотых стандартов — открытых для использования размеченных корпусов.

Программный комитет конференции выражает особую признательность Алексею Сорокину, Наталье Лукашевич, Анатолию Старостину и Виктору Бочарову, внесшим особый вклад в организацию и проведение этих тестирований.

Среди особых направлений «Диалога» в этом году — модели дискурса. Стало уже более или менее очевидно, что ключ к решению многих задач NLP — выход в лингвистическом анализе за границы отдельного предложения. В сборнике представлены работы, посвященные таким явлениям, как референциальный выбор, риторические отношения (RST) и т. п.

Как обычно, важную часть программы конференции и сборника составляют работы по анализу звучащей речи и, шире, проблемам мультимодальности — всего того, что подпадает под определение важного и традиционного для «Диалога» направления «Модели общения». Одной из самых ярких личностей в этой области на «Диалоге» и в российской лингвистике в целом была Елена Александровна Гришина. Ее работа была принята на «Диалог 2016» первой. К великому сожалению, публикация в этом сборнике будет последней: когда сборник уже верстался, пришла горестная весть о том, что Лены больше нет с нами.

Статьи в сборнике публикуются на русском и английском языках. При выборе языка публикации действует следующее правило:

- доклады по компьютерной лингвистике должны подаваться на английском языке. Это расширяет их аудиторию и позволяет привлекать к рецензированию международных экспертов.
- доклады, посвященные лингвистическому анализу русского языка, предполагающие знание этого языка у читателя, подаются на русском языке (с обязательной аннотацией на английском).

Несмотря на традиционную широту тематики представленных на конференции и отобранных в сборник докладов они не могут дать полной картины направлений «Диалога». Ее можно получить с помощью сайта конференции www.dialog-21.ru, на котором представлены обширные электронные архивы «Диалогов» последних лет и все результаты проведенных тестирований DialogueEvaluation.

Мы обращаем внимание авторов и читателей сборника, что его бумажный вариант, который вы держите в руках, является вторичным по отношению к сборнику, который размещается на сайте конференции и индексируется SCOPUS. Мы рекомендуем при цитировании использовать именно сетевую версию.

Программный комитет конференции «Диалог»
Редсовет сборника «Компьютерная лингвистика
и интеллектуальные технологии»

Организаторы

Ежегодная конференция «Диалог» проводится под патронажем Российского фонда фундаментальных исследований при организационной поддержке компании ABBYY.

Учредителями конференции являются:

- Институт лингвистики РГГУ
- Институт проблем информатики РАН
- Институт проблем передачи информации РАН
- Компания ABBYY
- Компания Yandex
- Филологический факультет МГУ

Конференция проводится при поддержке Российской ассоциации искусственного интеллекта.

Международный программный комитет

Байтин Алексей Владимирович	Компания Yandex, Россия
Богуславский Игорь Михайлович	Институт проблем передачи информации РАН им. А. А. Харкевича, Россия
Буате Кристиан	Университет Джозефа Фурье, Гренобль 1, Франция
Гельбух Александр Феликсович	Национальный политехнический институт, Мексика
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН им. А. А. Харкевича, Россия
Кобозева Ирина Михайловна	Московский государственный университет им. М.В. Ломоносова, Россия
Козеренко Елена Борисовна	Институт проблем информатики РАН, Россия
Корбетт Гревил	Университет Суррея, Великобритания
Кронгауз Максим Анисимович	НИУ «Высшая школа экономики», Россия
Лукашевич Наталья Валентиновна	НИВЦ МГУ им. М. В. Ломоносова, Россия
Маккарти Диана	Кембриджский университет, Великобритания
Мельчук Игорь Александрович	Монреальский университет, Канада
Нивре Йоаким	Уппсальский университет, Швеция
Ниренбург Сергей	Университет Мэриленда, Балтимор, США
Осипов Геннадий Семёнович	Институт программных систем РАН, Россия
Раскин Виктор	Университет Пердью, США
Селегей Владимир Павлович	Компания ABBYY, Россия
Хови Эдуард	Университет Карнеги — Меллон, США
Шаров Сергей Александрович	Университет Лидса, Великобритания

Организационный комитет

Селегей Владимир Павлович, *председатель*	Компания ABBYY
Байтин Алексей Владимирович	Компания Yandex
Беликов Владимир Иванович	Институт русского языка им. В. В. Виноградова РАН
Браславский Павел Исаакович	Уральский федеральный университет
Добров Борис Викторович	НИВЦ МГУ им. М. В. Ломоносова
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН им. А. А. Харкевича
Кобозева Ирина Михайловна	Московский государственный университет им. М. В. Ломоносова
Козеренко Елена Борисовна	Институт проблем информатики РАН
Лауфер Наталия Исаевна	Компания Yandex
Ляшевская Ольга Николаевна	Институт русского языка им. В. В. Виноградова РАН
Соколова Елена Григорьевна	Российский государственный гуманитарный университет
Толдова Светлана Юрьевна	НИУ «Высшая школа экономики»
Шаров Сергей Александрович	Университет Лидса

Секретариат

Атясова Анастасия Леонидовна, *координатор оргкомитета*	Компания ABBYY
Белкина Александра Андреевна, *секретарь оргкомитета*	Компания ABBYY
Гусева Анна Александровна, *координатор Dialogue Evaluation*	Компания ABBYY
Севергина Екатерина Александровна, *администратор оргкомитета*	Компания ABBYY

Рецензенты

Августинова Таня
Азарова Ирина Владимировна
Андрианов Андрей Иванович
Апресян Валентина Юрьевна
Байтин Алексей Владимирович
Баранов Анатолий Николаевич
Беликов Владимир Иванович
Бенко Владимир
Богданов Алексей Владимирович
Богданова-Бегларян Наталья Викторовна
Богуславский Игорь Михайлович
Бонч-Осмоловская Анастасия Александровна
Бочаров Виктор Владиславович
Браславский Павел Исаакович
Васильев Виталий Геннадьевич
Галицкий Борис Александрович
Гельбух Александр Феликсович
Гращенков Павел Валерьевич
Гришина Елена Александровна
Губин Максим Вадимович
Даниэль Михаил Александрович
Добров Борис Викторович
Добровольский Дмитрий Олегович
Зализняк Анна Андреевна
Захаров Виктор Павлович
Захаров Леонид Михайлович
Иомдин Борис Леонидович
Иомдин Леонид Лейбович
Катинская Анисья Юрьевна
Кибрик Андрей Александрович
Кобозева Ирина Михайловна
Коротаев Николай Алексеевич
Котельников Евгений Вячеславович
Котов Артемий Александрович
Крейдлин Григорий Ефимович
Кронгауз Максим Анисимович
Левонтина Ирина Борисовна
Лобанов Борис Мефодьевич
Лукашевич Наталья Валентиновна
Лютикова Екатерина Анатольевна
Маккарти Диана
Минлос Филипп Робертович
Наков Преслав
Недолужко Анна Юрьевна
Падучева Елена Викторовна
Пазельская Анна Германовна
Панченко Александр Иванович
Паперно Денис Аронович
Пиперски Александр Чедович
Плунгян Владимир Александрович
Подлесская Вера Исааковна
Рахилина Екатерина Владимировна
Селегей Владимир Павлович
Смирнов Иван Валентинович
Соколова Елена Григорьевна
Сомин Антон Александрович
Сорокин Алексей Андреевич
Старостин Анатолий Сергеевич
Степанова Мария Евгеньевна
Тестелец Яков Георгиевич
Тихомиров Илья Александрович
Толдова Светлана Юрьевна
Урысон Елена Владимировна
Федорова Ольга Викторовна
Хорошевский Владимир Федорович
Циммерлинг Антон Владимирович
Шаров Сергей Александрович
Янко Татьяна Евгеньевна

Содержание*

Приглашенные доклады

Основная программа конференции

---

* Доклады упорядочены по фамилии первого автора в соответствии с порядком английского алфавита | The reports of each section are ordered by the surname of the first author in compliance with the English alphabet.

Приглашенные доклады

DEEP LEARNING AND STRUCTURAL KERNELS FOR SEMANTIC INFERENCE: QUESTION ANSWERING APPLICATIONS TO FORMAL TEXT AND WEB FORUMS

Alessandro Moschitti (amoschitti@gmail.com)

Qatar Computing Research Institute, Hamad Bin Khalifa University, Qatar

In very recent years, there has been a large body of research work dedicated to deep learning: many recent articles report state-of-the-art results obtained by neural networks on many different applications of Natural Language Processing (NLP), e.g., Machine Translation, Speech processing, etc.

However, even the most optimist supporters of Neural Networks acknowledge the fact that dealing with highly complex semantic tasks requires new deep learning solutions. For example, no effective neural network model has been proposed so far for accurately performing discourse and dialog parsing or deep semantic inference, which is, for example, needed in Question Answering, Textual Entailment and Paraphrasing Identification.

The main difficulty in designing effective Neural Networks for such NLP tasks is the automatic learning of embeddings that can encode complex semantic relations, e.g., whose arguments (i) can span more than few consecutive words and (ii) can be located at any arbitrary distance in the target text. In other words, automatically learning semantic structures for language inference is still an open problem.

In this talk, I will elaborate on the claim above by firstly introducing properties and challenges of automatic language inference as well as the latest developments of the related advanced technology. The latter also includes the use of structural kernels applied to syntactic and semantic structures, which can easily encode complex text dependencies in learning algorithms. For this purpose, I will also capitalize on my direct experience with the techniques and models used for engineering the famous IBM Watson system.

Then, I will introduce some simple networks, along with some of their successful applications to simple NLP problems, e.g., sentiment analysis or named entity recognition, highlighting the importance of pre-training the networks with unsupervised models, e.g., using the famous toolkits, word2vec and GloVe.

Finally, I will present deep learning models for solving complex inference on short text, i.e., the one required for solving Question Answering (QA) applications. In particular, I will focus on Community QA, which offers the possibility of (i) utilizing a strict semantic correlation between questions and answers, i.e., user comments; (ii) modeling the user interaction structure, which is implicitly part of the question-comment treads; and (iii) dealing with a real-world application, which also requires the use of robust methods to process noisy text.

The results of our comprehensive comparative study on the above-mentioned machine learning methods suggest that the combination of deep learning and structural kernels applied to syntactic/semantic representations achieves the state of the art in applications requiring highly complex semantic inference. Therefore, although deep learning is greatly contributing to the solution of several NLP tasks, it is still important to combine it with traditional approaches, e.g., kernel methods applied to syntactic/semantic structure.

A THEORY OF CONTENT FOR NLP

Mark Steedman (steedman@inf.ed.ac.uk)

University of Edinburgh, UK

Linguists and computational linguists have come up with some quite useful theories of the semantics of function words and the corresponding logical elements such as generalized quantifiers and negation (Woods 1968; Montague, 1973; Steedman 2012). There has been much less progress in defining a usable semantics for content words.

The effects of this deficiency are very bad: linguists find themselves in the embarrassing position of saying that the meaning of „seek“ is seek‘. Computation lists find that their wide coverage parsers, which are now fast and robust enough to parse billions of words of web text, have very low precision as question answerers because, while the answers to questions like „Who wrote ‘The Great Gatsby?’“ are out there on the web, they are not stated in the form suggested by the question, „X wrote ‘The Great Gatsby’“, but in some other form that paraphrases or entails the answer, such as „X’s ‘The Great Gatsby’“.

Semantics as we know it is not provided in a form that supports practical inference over the variety of expression we see in real text.

I‘ll discuss recent work with Mike Lewis which seeks to define a novel form of semantics for content words using semi-supervised machine learning methods over unlabeled text. True paraphrases are represented by the same semantic constant. Common-sense entailment is represented directly in the lexicon, rather than delegated to meaning postulates and theorem-proving. The method can be applied cross-linguistically, in support of machine translation. If I have time, I will discuss the relation of this representation of content to the prelinguistic language of mind that must underlie all natural language semantics, but which has so far proved resistant to discovery.

1. *Lewis and Steedman,* (2013) „Combined Distributional and Logical Semantics“, Transactions of the Association for Comptuational Linguistics, 1, 179–192.
2. *Lewis and Steedman,* (2014) „Combining Formal and Distributional Models of Temporal and Intensional Semantics“, Proceedings of the ACL Workshop on Semantic Parsing, Baltimore, 28–32.

CONCURRENT DISCOURSE RELATIONS

Bonnie Webber (bonnie@inf.ed.ac.uk)

University of Edinburgh, UK

The Penn Discourse Treebank 2.0 (PDTB2) was released to the public in 2008 and remains the world's largest corpus of manually annotated discourse relations—both relations that are signaled explicitly (e.g., by a coordinating or subordinating conjunction, or by a discourse adverbial or other construction) and relations that otherwise appear implicit. Work is progressing on augmenting the PDTB2 in three ways: (1) by annotating many more forms of sentence-internal discourse relations; (2) by annotating implicit relations across paragraph boundaries; and (3) by identifying and annotating the presense of concurrent discourse relations. The new corpus will be called the PDTB 3.0 (or PDTB3).

The Penn Discourse TreeBank differs from other discourse-annotated corpora, not just in its size or its grounding in lexical, syntactic and/or positional evidence, but also in permitting more than one discourse relation to be annotated as holding concurrently.

In the PDTB2, annotators could indicate concurrent discourse relations by assigning multiple sense labels to an explicit connective. In the absense of an explicit connective, annotators could indicate concurrent discourse relations either by annotating a single implicit connective that they took to concurrently convey multiple senses or by annotating multiple implicit connectives, each conveying one of the concurrent relation(s). In the PDTB3, we have also had to allow for the possibility that a distinct implicit discourse relation can be inferred alongside an explictly signalled one.

Evidence for such concurrent relations comes from linguist-generated minimal pairs, from existing corpora, and from judgments elicited through crowdsourcing experiments that we have been carrying out for the past year.

There are different circumstances in which different concurrent discourse relations are taken to hold. I will go through these, and conclude with what I take the implications of this to be for various language technologies.

Основная программа конференции

THE IMPACT OF DIFFERENT DATA SOURCES ON FINDING AND RANKING SYNONYMS FOR A LARGE-SCALE VOCABULARY

Antonova A. (antonova@yandex-team.ru),
Kobernik T. (kobernik@yandex-team.ru),
Misyurev A. (misyurev@yandex-team.ru)

Yandex, Moscow, Russia

In this paper we compare different models for measuring synonymy. We consider methods based on monolingual text corpora and parallel texts. We experiment with the features based on context similarity, translation similarity, and similarity of neighbors in the parse trees. We provide an analysis of strong and weak points of different approaches and show that their combination can improve the results. The considered methods can handle large-scale vocabularies and be useful for automatic construction of human-oriented synonym dictionaries.

Keywords: synonyms, semantic similarity, synonym dictionary extraction, vector models, translation models

1. Introduction

In this paper we compare different models for measuring synonymy. Such methods can be useful for automatic construction of synonym dictionaries (see Fig. 1).

We consider methods based either on plain monolingual texts, or on parallel texts. Such corpora can be gathered from the Web and updated regularly with the growing number of documents. The automatically constructed synonym dictionaries can have the following advantages: coverage, variety, sensitivity to real occurrence in texts, and recent language changes. For this reason, it is interesting to compare automatically extracted synonyms and the synonyms from a human-built dictionary.

Many existing methods of automatic extraction of synonyms are based on the similarity of contexts in monolingual texts (see 2.2). However, many of the reported experiments are of a preliminary nature. The dependence between the quality of the results and the amount of data and the word frequencies is not quite understood. Context-based methods have some limitations, since the context similarity does not always directly correspond to synonymy. Problems can arise in the case of polysemous words. It is also difficult to collect reliable contextual information for rare words.

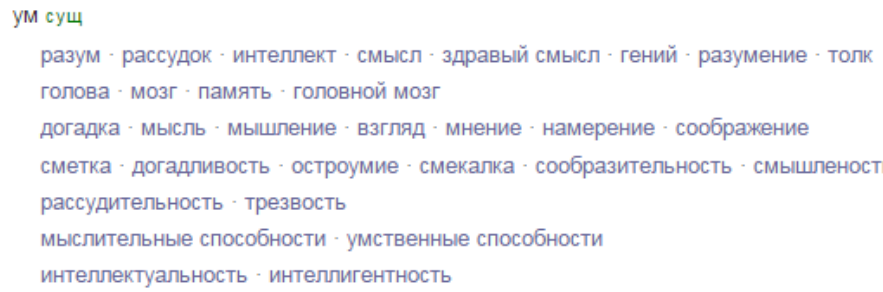


Fig. 1. Example of synonym dictionary entry for word «ум» ("mind, intellingence")

Another useful source of information about synonymy can be found in parallel corpora. Statistical translation models (phrase-tables) can serve as a source of synonym candidate pairs. One can assume that if two words have common translations in another language, they can be synonyms. Translation frequencies in the phrase-tables can be used to estimate the distance between synonym candidates.

It seems that having the information about the syntactic relations between words can also be useful [Dekang Lin 1997]. One could use the same context-based methods, but replace context words with syntactically related words, i.e. words adjacent in the parse tree. Such data is likely to contain less noise.

The aim of the paper is to study the impact of different data sources (namely, vector models, translation tables, syntax) on the quality of finding and ranking synonyms. We are interested in obtaining large and accurate human-oriented dictionaries. We are also interested in studying the dependence of the results on the size of the corpora, the size of vocabulary and word frequencies. We report on the experiments with combinations of different data sources for Russian.

The notion of synonymy is rather vague and subjective, which makes it difficult to find a reliable formal metric. The evaluation metrics based on assessments of human experts can be hard to reproduce. It is not clear how such metrics take into account the specifics of possible areas of practical application. For the evaluation of automatically extracted synonyms, we compare them to the synonyms from human-built dictionaries.

We restrict the pairs of words that can be considered candidates for synonyms, to the words that have at least one common translation in an SMT phrase-table. It turns out that part of manual synonym pairs cannot be found in this way [see Section 4.3]. On the other hand, the automatically extracted synonyms can be more relevant and up-to-date, since they reflect the word usage in real texts.

We regard as reference all synonym candidate pairs that intersect with human-built dictionary. The rest can contain both correct and not correct synonym pairs. A significant portion of automatically generated correct synonym pairs may not appear in the reference. In this approach, the task of finding the right synonym pairs is similar to the problem of ranking. Different models of similarity estimation produce the ranked lists of candidate pairs, which then can be compared w.r.t. the ranks of reference pairs. We believe that this experimental setup may be useful for the evaluation and analysis of different

methods. On the one hand, we rely on publicly available dictionary and do not need any further human assessment. On the other hand, the task has clear practical value.

The rest of the paper is organized as follows. In Section 2 we outline the related work. In Section 3 we describe the models of similarity estimation, and the features that we use. In Section 4 the experimental setup is described. The results of the experiments are reported in Section 5. We conclude and discuss the applicability of overall approach to building large-scale synonym dictionaries in Section 6.

2. Related work

The task of synonym extraction is closely related to the more general problem of measuring semantic similarity of words and phrases. The existing approaches can be roughly subdivided into three types according to the main source of information about semantic similarity.

2.1. Knowledge-based approaches

Knowledge-based methods try to make use of existing lexical resources, such as thesauri or knowledge graphs that represent a kind of a semantic network. The similarity of two words can be estimated, taking into account the distance between them in this network. Many approaches[Richardson 1994, Postma 2014] are based on Wordnet[Miller 1995, Fenenbaum 1998], a manually created lexical resource for English, but there also exist attempts to automatically induce semantic networks, e.g. from Wikipedia. However, such methods are left beyond the scope of this paper, since they require large-scale resources to be available for a particular language.

2.2. Monolingual context-based approaches

There exist different methods of measuring semantic similarity between two words based on the lexical coocurrence in large monolingual corpora. A wide variety of measures were proposed [see Baroni 2014 for systematic comparison] from simple scalar product of coocurrence frequency vectors, or Kullback-Leibler distance between the context distributions to more complex methods, that overcome the data sparsity problem, such as Latent Semantic Analysis (LSA) [Landauer and Dumais 1997], Distributional Memory [Baroni 2009] and neural network language models [Mikolov 2013, Pennington 2014].

The application of different methods to Russian is discussed in detail in the materials of Russe-2015 [Panchenko et al. 2015] contest. Different approaches are presented, including distributional, and neural network-based models, trained on a wide variety of monolingual corpora, as well as knowledge-based approaches.

2.3. Translation-based approaches

There exist methods for the extraction of synonym candidate pairs from bilingual parallel corpora. They use the alignment techniques from phrase-based statistical machine translation, and identify candidate synonyms using a phrase in another language as a pivot. [Dolan 2004, Bannard 2005, Barzilay 2001, Zhao 2008, Bansal 2012].

3. Models for similarity estimation

3.1. Translation model and extraction of synonym candidate pairs

A translation model $TM = (en_i, ru_j, count)$ is the set of translation equivalents with their respective counts. The translation equivalents are extracted from a parallel corpus with the help of SMT techniques and tools [Koehn 2003]. We also preprocess parallel sentences by a morpho-syntactic analyzer [Antonova, Misyurev 2012]. This allows us to sum over the counts the translations with the same lemmas, selecting only one pair in the translation model as described in [Antonova, Misyurev 2014].

The set of Russian synonym candidate pairs can be defined as

$$Cand(t) = \{(ru_a, ru_b) \mid \exists en_i, c_a, c_b, (en_i, ru_a, c_a) \in$$

$$\in TM, (en_i, ru_b, c_b) \in TM, c_a \geq t, c_b \geq t\} \quad (1)$$

Eq. 1 describes the set of pairs of Russian words for which there exists at least one common translation with a joint count above a given threshold t.

3.2. Translation similarity score

The similarity estimate for a pair of synonym candidate pair is calculated by Eq. 2.

$$TranslationSimilarity(ru_a, ru_b) = TranslationSimilarity(ru_b, ru_a) =$$

$$= Pr(ru_a|ru_b)Pr(ru_b|ru_a) \quad (2)$$

$$Pr(ru_a|ru_b) = \sum_i Pr(ru_a|en_i)Pr(en_i|ru_b) \quad (3)$$

$$Pr(ru_b|ru_a) = \sum_i Pr(ru_b|en_i)Pr(en_i|ru_a) \quad (4)$$

$$Pr(en_i|ru_a) = \frac{Count(en_i, ru_a)}{\sum_{en_i} Count(en_i, ru_a)} \quad (5)$$

For all common English translations en_i we calculate the probabilities of translating ru_a to ru_b through en_i, and then marginalize over en_i to get $Pr(ru_a|ru_b)$. To get a single symmetrical similarity estimate we get a multiplication of $Pr(ru_a|ru_b)$ and $Pr(ru_b|ru_a)$.

Typical mistakes observed when ranking synonym candidate pairs by translation similarity score are the following:

- Mistakes of automatic word alignment may produce incorrect translations pairs. Particularly, wrong pieces of multiword expressions can be found in the phrase table with high counts.
- Mistakes introduced by the word sense ambiguity. For example, unrelated Russian words «бежать» (“move quickly”) and «запускать» (“launch”) can be translated by a polysemous English word “run”. Such polysemous common translations can connect words that are not synonyms.

These two types of mistakes are specific for the phrase-table and one can expect that a combined approach taking monolingual contexts into account can improve the ranking. It is important to use a lemmatized phrase-table, otherwise the synonym candidate pairs can contain many word forms of the same lemma and the counts can be much sparser.

3.3. Similarity scores by vector models

For measuring similarity in monolingual contexts we train vector models with the help of two popular tools *word2vec* and *glove*. They represent each word as a vector in a low-dimensional space. The similarity score between two words is given by the scalar product of the corresponding vectors. Though vector models are widely popular and effective, this approach still has some weaknesses:

- It cannot divide separate meanings of polysemous words, since each word has only one corresponding vector.
- Vectors for rare words can be unreliable, since they occur in a small number of contexts in the corpus.
- It is appropriate for single words, but requires special efforts to handle multiword expressions.

3.4. Vector models with syntactically related words

We checked the possibility of using syntactic relations to construct vector models of words. As is known, such models are based on the co-occurrence statistics of the corpus. Instead of collecting all words within a predefined window, we collected syntactic contexts, i.e. the words that are adjacent in the parse tree. Such data is likely to contain less noise. We trained a *glove* vector model on the set of syntactic contexts. In the rest of the paper this model is referred to as *GloveSynt*.

3.5. Model combination

It seems that context similarity and translation similarity suffer from different types of problems and that their combination can improve the results. Moreover, taking word frequencies into account can possibly lead to further improvement. We combine the similarity scores by different models and word frequencies in a log-linear model, and train the weights with logistic regression.

3.6. Quality metrics

The notion of synonymy is rather vague and subjective, which makes it difficult to find a reliable formal metric. To assess the quality of a ranked list of automatically extracted synonyms, we look at the ranks of the gold synonyms from a human-built dictionary. We evaluate the importance of different features w.r.t. the ranking that they produce on the list of all candidate pairs. To measure the ranking quality we use the following metrics: average precision (*AveP*), average rank(*AveRank*), median of ranks(*Median*). The average precision is defined as follows:

$$AveP = \frac{\sum_{r=1}^{n} P(r) \times rel(r)}{\sum_{r=1}^{n} rel(r)} \quad (6)$$

where r is the rank in the sequence of candidates, $P(r)$ is the precision of top r candidates, $rel(r)$ is an indicator function equaling 1 if r-th pair is relevant, n is the total number of candidates.

4. Experiment

4.1. Reference synonym pairs

We downloaded Russian synonym dictionary from Wiktionary.org, taking only semantic relation of type "Synonym". The initial 58,715 synonym pairs were lowercased, and symmetrized[1] (105,142 pairs after symmetrization). Considering only single-word pairs for the described experiment we got a set of reference pairs, Gold = {(query, synonym)}, which contained 99,394 single-word synonym pairs for 42,509 distinct queries.

Another dataset *GoldAbr* consisted of Abramov's dictionary of Russian synonyms and similar words, whose first edition was in 1915. After symmetrization it contained 34,930 pairs for 12,527 distinct queries. The intersection of Wiktionary and Abramov dictionary is small: 6,616 pairs.

[1] Though some synonym pairs (a, b) may be asymmetrical, most added pairs (b, a) are also relevant.

4.2. Synonym candidate pairs

We built a lemmatized phrase-table with maximum phrase length of 3 words on an English-to-Russian corpus drawn from the Web. The minimal joint translation count is 2. The sum of all joint counts is about 2.91 billion. For the simplicity of the experiment, we restricted the Russian side only to single words that had been recognized as correct Russian lemmas by an in-house morphological dictionary.

We generated a set of synonym candidate pairs with the help of the phrase-table, as described in 3.1. The set of positive examples consisted of the intersection of reference and candidate pairs $Pos = Gold \cap Cand$. The set of negative examples consisted of those pairs, whose query word occurred in *Gold* and had at least one positive example in *Pos*.

All query words were randomly divided into two parts. Then all positive and negative examples were place into one of two sets according to their query word:

- training set: 26,281 positive, 2,657,507 negative examples.
- test set: 26,461 positive, 2,614,819 negative examples.

Fig. 2 represents the dependence between the number of candidate synonyms and the query frequency.

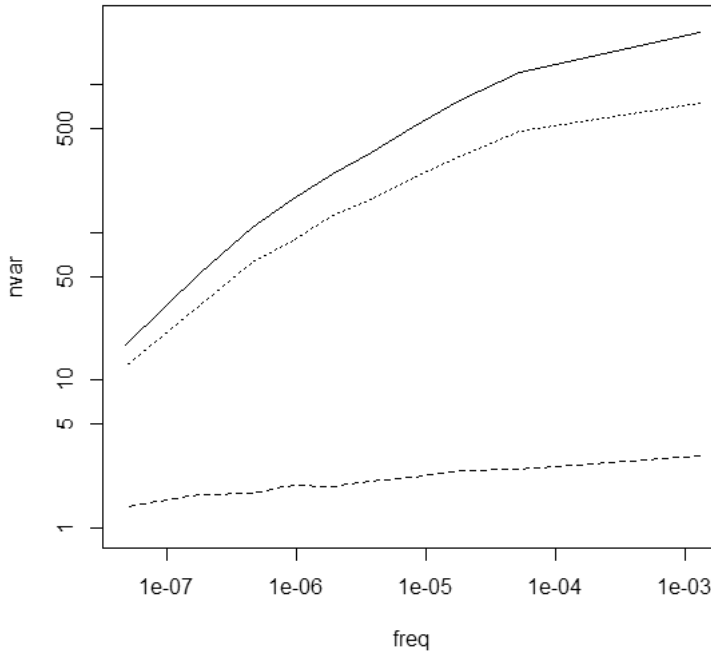


Fig. 2. Dependence between the number of candidate synonyms and the query frequency. Bottom line — reference synonyms, upper line — all candidates by translation model, the middle line — number of candidates per one reference pair

Frequent words typically have more synonyms in human-built dictionaries, and more synonym candidate pairs. The absolute number of candidate pairs is very big (logarithmic scale). On the one hand we have more data for more frequent words. On the other hand the classification task is harder because they have more synonym candidates.

Concerning the task of synonym evaluation, there exists a problem of reference sparsity. It means that the candidate list often contains many relevant synonym pairs that are missing in the reference (see Table 1). For that reason, the use of standard metrics such as recall/precision may be unconvincing.

Table 1. Top-34 synonym candidates for word «проворный» ("agile"), ranked by descending translation similarity. Reference synonyms are bold (Wiktionary) and underlined (Abramov)

верткий	(~nimble)	$1.8e^{-3}$	**бойкий**	(~spirited)	$7.6e^{-6}$
ловкий	(~agile)	$9.4e^{-4}$	незамедлительный	(~immediate)	$7.2e^{-6}$
поворотливый	(~agile)	$8.7e^{-4}$	сноровистый	(~nimble)	$5.2e^{-6}$
шустрый	(~nimble)	$6.6e^{-4}$	подвижной	(~mobile)	$5.0e^{-6}$
юркий	(~nimble)	$2.7e^{-4}$	подсказывающий	(~prompting)	$3.4e^{-6}$
прыткий	(~nimble)	$1.7e^{-4}$	динамичный	(~dynamic)	$2.7e^{-6}$
маневренный	(~maneuvering)	$9.6e^{-5}$	оживленный	(~brisk)	$2.6e^{-6}$
быстрый	(~fast)	$5.4e^{-5}$	своевременный	(~timely)	$2.3e^{-6}$
быстро	(~quickly)	$5.3e^{-5}$	находчивый	(~resourceful)	$1.7e^{-6}$
расторопный	(~agile)	$5.2e^{-5}$	изворотливый	(~quirky)	$1.6e^{-6}$
гибкий	(~flexible)	$4.7e^{-5}$	безотлагательный	(~urgency)	$1.4e^{-6}$
вертлявый	(~fidgety)	$3.6e^{-5}$	пробужденный	(~awakened)	$8.2e^{-7}$
подвижный	(~mobile)	$3.1e^{-5}$	активный	(~agile)	$8.1e^{-7}$
оперативный	(~operational)	$1.8e^{-5}$	скорый	(~fast)	$8.0e^{-7}$
стремительный	(~rapid)	$1.7e^{-5}$	сообразительный	(~witted)	$8.0e^{-7}$
скорейший	(~early)	$1.4e^{-5}$	**резвый**	(~spirited)	$6.3e^{-7}$
подсказанный	(~prompted)	$8.8e^{-6}$	бодрый	(~brisk)	$6.2e^{-7}$

Fig. 3 illustrates the problem of measuring the ranking quality with the average precision when the reference is sparse. One can see that although some reference pair are ranked high according to our similarity models, a considerable amount of reference pairs belong to the low-precision area. For that reason we used additional metrics, namely, average rank(*AveRank*), median of ranks(*Median*).

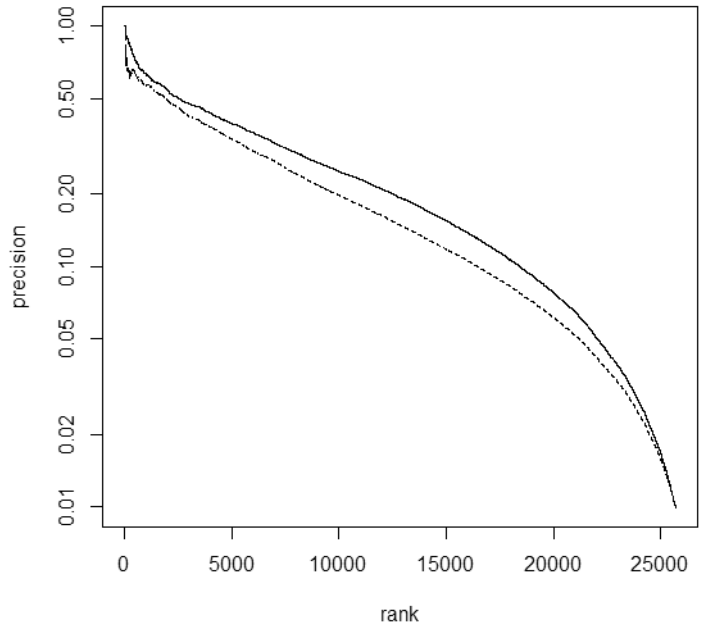


Fig. 3. Precision vs rank. Solid line — translation similarity, dashed line — word2vec

4.3. Missing synonym pairs

Only 58.9% of initial reference pairs were found among candidates generated by the phrase-table. Among those missing pairs, 61.3% are the pairs in which one or both words had not occurred in the phrase-table at all, 38.7% are the pairs in which both words occurred in the phrase-table, but had no common translations.

We manually annotated 100 random missing pairs for Wiktionary (see Table 2). Only 36% of them were actually good synonyms, though they also included words with low frequencies.

Table 2. Human annotation results for 100 random reference pairs without common translations

%	Human judgement	Example
36	Good candidates	ряженка — варенец (~milk product), пацанва — детвора (~children)
25	Both words are rare or unknown	тонемика — тонология (~tonology?)
21	One word is rare or unknown	гуртоправ — гуртовщик (~drover), скворец — кокако (~starling)
6	One word is slang or obscene	записка — малявка (~note)
5	Words are not synonyms	важный (~important: adjective) — царственно (~kingly: adverb)
4	One word is not Russian	галоген — галоїд (~halogen)
3	Obsolete meaning	сплетник (~gossip) — трубач (~trumpeter), управлять (~to control) — рядить (~to dress?)

4.4. Features for model combination

We calculated the following features for each synonym candidate pair from test set and training set.

1. Logarithm of translation similarity score (see Eq.2).
2. Scalar product of vectors by the word2vec model. The model was trained with standard parameters, the vocabulary consisted of words occurring at least 10 times in the corpus. The corpus consisted of 200 mln sentences, disambiguated and lemmatized with a morpho-syntactic analyzer [Antonova, Misyurev 2012].
3. Scalar product of vectors by the glove model. The training setting was the same as previous.
4. Logarithms of the frequencies of the two synonyms in the monolingual corpus.
5. Scalar product of vectors by the glove model trained on syntactic contexts. The contexts include lemmas that are adjacent to the given word in the parse trees.

We combined the combinations of the above features in a log-linear model, and trained the weights with logistic regression.

5. Experiment results

We report the ranking quality given by different models and their combinations. A summary of these can be seen in Table 3.

Table 3. Ranking results for different feature combinations. The metrics are average precision (*AveP*), average rank (*AveRank*) and median of ranks (*Median*)

Feature combination	AveP	AveRank	Median
Wiktionary dataset			
Word2Vec	0.165	407,228	132,683
TranslationSimilarity	0.237	222,513	66,802
TranslationSimilarity + Frequencies	0.247	212,819	65,304
Word2Vec + TranslationSimilarity + Frequencies	**0.303**	**181,381**	**50,232**
Glove	0.117	560,219	219,061
Glove + TranslationSimilarity + Frequencies	0.299	182,327	50,338
GloveSynt	0.058	803,457	467,868
GloveSynt + TranslationSimilarity	0.274	204,993	57,393
GloveSynt + TranslationSimilarity + Frequencies	0.291	191,225	54,492
Abramov dataset			
Word2Vec	0.025	516,313	244,381
Word2Vec + TranslationSimilarity + Frequencies	0.068	250,096	79,268
Glove	0.031	506,889	220,102
Glove + TranslationSimilarity + Frequencies	**0.075**	**238,683**	**73,224**
TranslationSimilarity	0.049	272,115	99,969

The vector model with syntactic contexts (*GloveSynt*) yields an improvement in combination with translation similarity model. However, the vector models trained on simple lemmatized text yield better results. Besides, using syntactic contexts requires parsing the corpus, which makes the experiments more complex, time-consuming and difficult to reproduce.

The classification results with the single *glove* model turned to be lower than those of *word2vec* model for Wiktionary dataset, but higher for Abramov dataset. It is interesting that in combination with the translation similarity model, they are almost equal in quality. The advantage of *glove* tool is that it allows us to parallelize the context extraction, e. g. with map-reduce operations.

It seems that Abramov dictionary contains less straightforward synonyms and more distant synonyms than Wiktionary. The absolute values of average precision for Abramov dataset is much lower than that for Wiktionary dataset.

Fig. 4 demonstrates the advantages of combination of different models for classification and ranking. Though they correlate on many examples, using two dimensions makes it possible to classify correctly some uncertain points.

Fig. 5 demonstrates the top-ranked pairs by different models depending on the word frequencies. One can see that the vector models tend to rank higher frequent words, while the translation similarity model ranks better rare words, but tends to prioritize words with similar frequency.

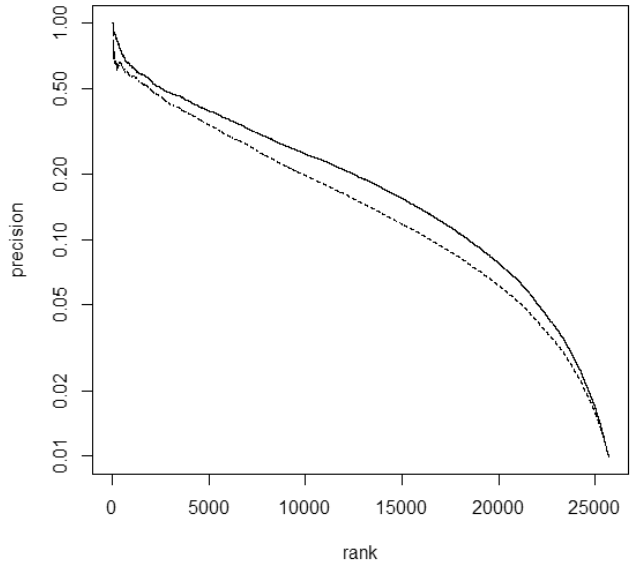


Fig. 4. Joint distribution of positive (○) and negative (×) examples w.r.t. TranslationSimilarity score and word2vec distance (Wiktionary dataset). Pt—is translation similarity score

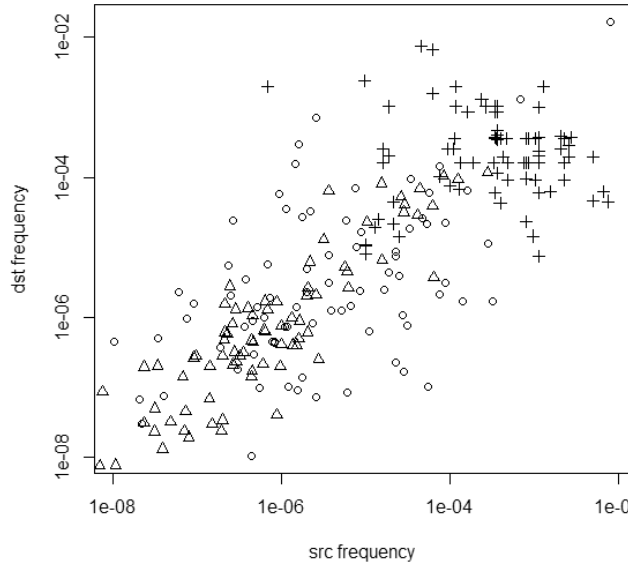


Fig. 5. Distribution of candidate pairs (src, dst) w.r.t. the word frequencies in Wiktionary dataset (○ — random reference pairs, Δ — top by translation similarity, × — top by vector models)

Conclusion

We compare the ranking of synonym candidate pairs given by different models. We consider models based on syntactic relations, monolingual contexts, and models based on parallel texts.

The set of synonym candidate pairs is generated with the help of the phrase-table, which is extracted from parallel texts with SMT methods. The translation similarity model based on the phrase-table statistics also proves to be useful for ranking candidates. It can handle rare and polysemous words.

We show that the precision of different models depends on word frequencies. Our experiments demonstrate that the combination of monolingual vector models and translation similarity model improves the ranking results, as well as taking word frequencies into account. The general approach has a practical value, since it can handle large-scale vocabularies and be useful for automatic construction of synonym dictionaries.

References

1. *Alexandra Antonova and Alexey Misyurev.* (2012), Russian dependency parser SyntAutom at the Dialogue-2012 parser evaluation task. Proceedings of the Dialogue-2012 International Conference.
2. *Alexandra Antonova and Alexey Misyurev.* (2014), Automatic creation of human-oriented translation dictionaries. Proceedings of the Dialogue-2014 International Conference.
3. *Bannard C., Callison-Burch C.* (2005), Paraphrasing with bilingual parallel corpora, Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, p. 597–604, June 25–30, 2005, Ann Arbor, Michigan.
4. *Bansal M., Denero J., Dekang Lin.* (2012), Unsupervised translation sense clustering, Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, June 03–08, Montreal, Canada.
5. *Baroni M., Dinu G., Kruszewski G.* (2014), Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Vol. 1), Baltimore, MD, USA, pp. 238–247.
6. *Baroni M., Lenci A.* (2009), One distributional memory, many semantic spaces. Proceedings of the EACL GEMS Workshop. Athens, Greece, pp. 1–8.
7. *Barzilay R., McKeown K. R.* (2001), Extracting paraphrases from a parallel corpus. ACL'01 Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, pp. 50–57
8. *Dekang Lin.* (1997), Using syntactic dependency as local context to resolve word sense ambiguity. In In Proceedings of ACL-97, page 64–71.
9. *Dolan W., Quirk C., Brockett C.* (2004), Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources, International Conference on Computational Linguistics, August 2004.

10. *Fellbaum C., ed.* (1998), WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA.
11. *Koehn P., Och F. J., Marcu D.* (2003), Statistical phrase-based translation. In Proceedings of NAACL.
12. *Landauer T. K., Dumais S. T.* (1997), A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge, Psychological Review, Vol. 104(2), pp. 211–240.
13. *Mikolov T., Chen K., Corrado G., Dean J.* (2013), Efficient Estimation of Word Representations in Vector Space.
14. *Miller G.* (1995), Wordnet: A lexical database for English. In Communications of the ACM.
15. *Panchenko, A., Loukachevitch, N. V., Ustalov, D., Paperno, D., Meyer, C. M., Konstantinova, N.* (2015), RUSSE: The First Workshop on Russian Semantic Similarity. In: Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue". Volume 2. RGGU, Moscow 89–105.
16. *Pennington J., Socher R., Manning, C. D.* (2014), Glove: Global vectors for word representation, Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP), Doha, Qatar, 1532–1543.
17. *Postma M., Vossen P.* (2014), What implementation and translation teach us: the case of semantic similarity measures in wordnets, Proceedings of Global WordNet Conference 2014, Tartu, Estonia, pp. 133–141.
18. *Richardson R., Smeaton A., Murphy J.* (1994), Using WordNet as a Knowledge Base for Measuring Semantic Similarity between Words, Proceedings of AICS Conference, Dublin, Ireland.
19. *Zhao S., Wang H., Liu T., Li S.* (2008), Pivot approach for extracting paraphrase patterns from bilingual corpora. In ACL: Annual Meet. of Assoc. of Computational Linguistics

ГЛАГОЛЫ *ИСЧЕЗНУТЬ* И *ПРОПАСТЬ*: МНОГОЗНАЧНОСТЬ И СЕМАНТИЧЕСКАЯ МОТИВАЦИЯ[1]

Апресян В. Ю. (valentina.apresjan@gmail.com)

Национальный исследовательский университет Высшая школа экономики; Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

В работе рассматриваются синонимичные глаголы *исчезнуть* и *пропасть* и анализируются различия в их семантике, мотивирующие их различия в синтаксических, аспектуальных и сочетаемостных свойствах, а также в структуре их полисемии. Семантические параметры, различающие эти глаголы, а именно, тип и денотативный статус объекта, причина прекращения существования, ожидания говорящего, скорость и полнота прекращения существования, наличие наблюдателя, представляются применимыми к анализу всего глагольного поля конца существования.

Ключевые слова: существование, конец, местонахождение, наблюдаемость, наблюдатель, полисемия, многозначность, синонимы, денотативный статус, валентность, аспектуальный, процессный, результативный, семантическая мотивация

ISCHEZNUT' 'TO DISAPPEAR' AND *PROPAST'* 'TO VANISH': POLYSEMY AND SEMANTIC MOTIVATION

Apresjan V. Ju. (valentina.apresjan@gmail.com)

National Research University Higher School of Economics; Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

The paper considers Russian synonymic verbs *ischeznut'* 'to disappear' and *propast'* 'to vanish' and analyzes their semantic differences which motivate their differences in syntactic, aspectual and collocational properties, as well as in their polysemies. Semantic oppositions that distinguish between these two verbs, namely, type and referential status of the disappearing object, cause of disappearance, speaker's expectations, speed and completeness

[1] Исследование выполнено за счет гранта Российского научного фонда (проект №16-18-02054, «Исследование русского языкового сознания на основе семантического, статистического и психолингвистического анализа лексической многозначности»).

of disappearance, presence of an observer can be applied to the analysis of the entire semantic domain of the 'end of existence'.

Key words: existence, end, location, observer, polysemy, synonyms, referential status, aspectual, process, result, semantic motivation

Введение

В работе представлено семантическое поле 'исчезновения', в первую очередь многозначные глаголы *исчезнуть* и *пропасть*, со- и противопоставление которых, насколько нам известно, еще не становилось предметом отдельного лингвистического исследования.

Представляется, что семантический анализ этих глаголов в их исходных значениях, в первую очередь, семантических компонентов 'тип объекта', 'причина прекращения существования', 'местонахождение', 'наблюдаемость' и 'ожидания говорящего', позволяет мотивировать их различия в структуре полисемии, а также в синтаксических, аспектуальных и прагматических свойствах. Кроме того, выделенные параметры представляются релевантными для всего поля конца существования.

1. Основные семантические противопоставления поля 'конца'

В работе (Кустова 2002:73–81) рассматриваются различия между основными фазово-бытийными глаголами (*начаться/кончиться*, *возникнуть/появиться/исчезнуть*) с точки зрения их сочетаемости с разными типами непредметных сущностей, например «темпоральными ситуациями», «информационными ситуациями», а также с точки зрения их собственной семантики — 'начало/конец процесса', 'начало/конец существования объекта'. Обобщая эти противопоставления, можно выделить следующие основные семантические типы внутри глаголов со значением 'конца':

- Прекращение собственных действий или деятельности субъекта: *перестать, прекратить, кончить, бросить, оставить, завязать*.
- Прекращение чужих действий или деятельности: *прекратить, покончить, прикрыть*.
- Прекращение процессов: *прекратиться, кончиться*.
- Прекращение существования объектов (материальных и нематериальных): *исчезнуть, пропасть*; *сгинуть*; *кануть*; *пасть*; *умереть*; *отмереть*; *вымереть*, *вывестись*; *улетучиться*, *испариться*, *растаять*; *сойти на нет*; *кончиться, иссякнуть, перевестись*; *уйти, пройти*.

Это поле чувствительно к различиям между разными типами предикатов [о фундаментальной классификации предикатов см. Апресян 2006:75–96],

т. е. отражает существующую в языке и имеющую многочисленные лингвистические проявления иерархию агентивности. Кроме того, оно пронизано и другими противопоставлениями, задаваемыми

- системообразующими смыслами — фундаментальными языковыми смыслами, организующими семантическую систему языка в целом [Апресян 2001], такими как причина, оценка, наблюдатель,
- а также противопоставлениями между разными таксономическими классами объектов (живые организмы, ситуации, ментальные состояния, ресурсы и пр.).

2. Семантика глаголов со значением ‘прекращения существования’

Основные глаголы ‘прекращения существования’ — это *исчезнуть, пропасть*; *сгинуть*; *кануть в вечность*; *пасть*; *умереть*; *отмереть*; *вымереть*, *вывестись*; *улетучиться*, *испариться*, *растаять*; *сойти на нет*; *кончиться, иссякнуть, перевестись*; *уйти, пройти*.

Главные семантические противопоставления в группе глаголов со значением ‘прекращения существования’ касаются

- типа объекта;
- денотативного статуса объекта;
- причины прекращения существования объекта;
- ожиданий говорящего;
- скорости и степени прекращения существования;
- наличия наблюдателя.

Все глаголы данной группы описывают прекращение существования какого-либо объекта. Некоторые из них имеют также другие значения, например, *исчезнуть* и *пропасть*, у которых есть значения ‘перестать находиться где-либо’ (*Со стола исчезли деньги*) и ‘перестать быть наблюдаемым’ (*Он исчез в тумане*), что является семантически мотивированным развитием значения.

В самом деле, каждый из этих семантических компонентов в норме прагматически имплицирует другие: если что-то наблюдаемо, то оно существует и где-то находится, если что-то существует, то оно наблюдаемо и где-то находится; если что-то где-то находится, то это существует и наблюдаемо и т. д. [ср. Падучева 2004: 110 о «правдоподобной прагматической инференции» компонента ‘переместиться’ из ‘перестать быть видимым’].

Таким образом, в полисемии некоторых глаголов со значением прекращения существования объекта объясняемым образом присутствуют также значения ‘изменение местонахождения’ и ‘утрата наблюдаемости’. Регулярная многозначность ‘находиться’, ‘существовать’, ‘быть видимым’ для глаголов появления и исчезновения и ее обсуждение в лингвистической литературе подробно разбирается в той же работе [Падучева 2004: 27, 110, 201, 242, 254, 439, 457, 481].

Однако именно данная структура многозначности является тенденцией, а не правилом, и у некоторых из приводимых здесь глаголов полисемия устроена иначе. Так, у глагола *кануть* в составе фраземы *как в воду кануть*, у глаголов *улетучиться* и *испариться* с исходным значением 'переход в другое состояние' есть значение 'переместиться', но не 'перестать быть видимым'; ср. *Он уже три дня как в воду канул*, *Он куда-то улетучился/испарился*, но не **Он канул в тумане*, **Он улетучился/испарился за дверью.*

У некоторых глаголов значение 'переместиться' является исходным, а значение 'прекращение существования' — производным (*пройти, уйти*).

У некоторых глаголов не развивается ни значение прекращения местонахождения, ни значение прекращения наблюдаемости; ср., например, глаголы *умереть, отмереть, вымереть, вывестись, кончиться*, *иссякнуть, перевестись*.

В данном разделе все глаголы анализируются в значении 'прекращение существования', которое для некоторых глаголов является основным (*умереть, вымереть, иссякнуть* и др.), для некоторых — одним из основных (*исчезнуть, пропасть*), а для некоторых — производным (*пасть, улетучиться*, *испариться*, *растаять* и др.).

2.1. Тип и денотативный статус объекта

Главное семантическое противопоставление в этой группе задается типом объекта, который перестает существовать; это может быть

- живой организм (*умереть, вымереть, вывестись*);
- часть живого организма (*отмереть*);
- ресурс, чаще материальный (*кончиться, иссякнуть* и особенно *перевестись*);
- крупное социальное образование (*пасть*; ср. *Империя/цивилизация пала*);
- состояние, например, *желание* (*улетучиться*, *испариться*, *растаять*; *уйти, пройти*).

При этом внутри каждой из групп есть более мелкие градации. Так, *умирать* могут отдельные живые организмы, а *вымирают* и *выводятся* — только группы, в случае *вымереть* — целые виды. С этим связан денотативный статус субъектов при этих глаголах — при *умереть* он может быть конкретно-референтным, при *вымирать* и *выводиться* — только родовым; ср. *Динозавры вымерли миллионы лет назад*; *В наших лесах полностью вывелись лоси*, но не **Все наши лошади вымерли <вывелись>*.

Кончиться может практически любой истощаемый ресурс, как материальный, так и не материальный: еда (*Молоко кончилось, Хлеб кончился*), одежда (*Чистые рубашки кончились*), другие материальные и не-материальные ресурсы (*Сигареты/боеприпасы кончились*, *Деньги кончились*, *Нефть кончилась*, *Бензин кончился*; *Работа кончилась*); физические и не-физические ресурсы человека (*Силы кончились*, *Терпение кончилось*).

Употребления с названиями разного рода положений дел и процессов (*Зима кончилась*, *Перемирие кончилось*, *Кровопролитие кончилось*) относятся к другому значению этого глагола, поскольку указывают не на прекращение существования ресурса в силу его исчерпанности, а на прекращение ситуации в силу достижения ее естественного временного завершения (*Зима кончилась*), либо в силу каких-то других причин (*После двух дней затишья перемирие кончилось*).

Иссякнуть могут либо материальные ресурсы, понимаемые как совокупность (*Запасы еды и воды быстро иссякли*, *Деньги иссякли*), но не как отдельные объекты (**Хлеб иссяк*, при нормальности *Запасы хлеба иссякли*); источники воды (*Родник иссяк, Фонтан иссяк*), причем не только природные (*Слезы иссякли*), а также не-физические ресурсы человека (*Фантазия иссякла, Энтузиазм иссяк, Терпение иссякло*).

Перевестись могут люди, обладающие особыми положительно оцениваемыми качествами, и поэтому воспринимаемые как некий полезный ресурс: *Перевелись богатыри на Руси*; *Романтики/энтузиасты нынче перевелись*.

Улетучиться, испариться и *растаять* описывают, как правило, внутренние, обычно эмоциональные, состояния: *веселость, злость, решимость*, *страх*, но не такие, которые имеют сильные внешние проявления (?*Ее паника испарилась/улетучилась/ растаяла*) или воспринимаются как труднопреодолимые (ср. странность ?*Его горе испарилось/улетучилось/растаяло*).

При этом *улетучиться* и *испариться*, в отличие от *растаять*, могут описывать не только ощущения самого субъекта, но и восприятие его внутреннего состояния наблюдателем; ср. *Его самодовольство на глазах куда-то улетучилось/испарилось*, при странности ?*Его самодовольство на глазах растаяло.*

Уйти и *пройти* могут как физические (*боль*), так и эмоциональные (*страх, гнев, обида, досада, тревога, влюбленность*) состояния, причем разной длительности и глубины (*любовь, горе* vs. *увлечение*, *раздражение*).

Оставшиеся глаголы могут, в той или иной степени, описывать прекращение существования разных типов объектов.

Наиболее широкой сочетаемостью обладает глагол *исчезать* — *исчезать* могут и живые организмы (*динозавры, лошади Пржевальского*), и социальные образования, как материальные, так и нематериальные (*цивилизации, народности, этносы, языки, города, обычаи*), а также ментальные, эмоциональные и физические состояния (*сомнения, страх, боль*), и некоторые виды физических объектов (*прыщи, пятна, бородавки, веснушки, морщины*).

При этом в применении к живым организмам *исчезнуть* может обозначать только прекращение существования вида, а не отдельного организма и, соответственно, употребляется только с родовыми субъектами: *Динозавры исчезли в конце мелового периода, возможно, под воздействием похолодания*, но не **Наши лошади исчезли под воздействием сапа*. Сочетание *Наши лошади исчезли*, естественно, возможно в значении ‘перестали находиться’ или ‘перестали быть видимыми’.

Однако не любой объект может *исчезать*: некоторые виды нематериальных объектов, раз возникнув, как бы вечно продолжают свое существование; ср. странность ?*Идея исчезла* [отмечается в работе Кустова 2002:75] и отсутствие подобных примеров в корпусе НКРЯ. Возможны при этом сочетания типа

Мысль исчезла; ср. *Мысль исчезла, как исчезает след капли на стекле* (Г. Щербакова). Различия между *идеей* и *мыслью* подробно анализируются в работе [Кобозева 1993], где *идея* рассматривается как разновидность мысли, а особый статус *идей* среди прочих мыслей как раз «проявляется в их относительной стабильности» [Кобозева 1993:103].

У *пропасть* сочетаемость существенно более узкая; в значении прекращения существования невозможны живые организмы и социальные образования (**Динозавры пропали в конце мелового периода*, **Прусский язык пропал в 17 веке*), но возможны состояния (*Желание пропало*) и некоторые виды физических объектов (*прыщи, бородавки* и пр.).

Сгинуть в значении прекращения существования обычно используется применительно к людям; ср. *Александр Аркадьевич пополнил уже длинный список российских учёных, сгинувших в последнее время от рук наших отморозков* («Криминальная хроника», 2003.07.08), но возможны и сочетания с названиями физических объектов *Доходы от ядерных отходов сгинули в Бермудском треугольнике* («Известия», 2002.01.21). Другое частотное употребление *сгинуть*, однако уже в значении 'переместиться' — применительно к разного рода нечистой силе, особенно в форме императива: *Сгинь, исчадие ада*, *Сгинь, нечистый*.

Кануть в вечность может любой материальный и нематериальный объект, который имеет какую-то социальную значимость и в силу этого хранится в памяти людей; его безвозвратный уход из этой общественной памяти и сознания и обозначается данной фраземой: *И канула в вечность фамилия его предков* (К. Кожевникова); *Отправивший в вечность сотни тысяч безвинных людей, он сам канул в вечность, перемолотый репрессивными жерновами* (Н. Прислонов).

Сойти на нет могут разные нематериальные объекты, имеющие степень или интенсивность (*поток инвестиций, уровень продаж*, *роман, любовь, желание, интерес*); ср. *Отношения их совсем сошли на нет* (С. Спивакова); *Веры в старом смысле нет, однако способность верить ещё не кончилась и не сошла на нет* (В. Маканин).

2.2. Причина прекращения существования и ожидания говорящего

По признаку причины прекращения существования глаголы делятся на следующие группы: те, для которых причина обязательна (*кончиться, иссякнуть*); те, у которых причины чаще нет (*пропасть, улетучиться, испариться*) и остальные, для которых причина в разной степени возможна, но не обязательна. С отсутствием или наличием причины прекращения существования связаны ожидания говорящего: глаголы, которые обычно указывают на беспричинность прекращения существования, одновременно указывают и на нарушение ожиданий говорящего.

По этому признаку противопоставлены близкие синонимы *исчезнуть* и *пропасть*: *исчезнуть* часто указывает на внешнее воздействие или целенаправленное усилие как фактор, приводящий к прекращению существования объекта, хотя можно и *таинственно/загадочно исчезнуть*.

У этого глагола есть валентность причины, которая выражается различными предложно-именными группами; ср. *исчезнуть под натиском цивилизации*; *исчезнуть из-за изменения климата*; *исчезать от нагревания*; *исчезать при стрессе*; *исчезнуть вследствие наводнения*. Ср. также примеры из НКРЯ: *Ядовитость [мышьяка] может исчезнуть от разложения* (Н. Амосов); *Старик медленно пожевал почти исчезнувшими от старости губами* (С. Бабаян).

Глагол *пропасть* часто указывает на беспричинное, с точки зрения говорящего, прекращение существования и на нарушение его ожиданий; ср. естественность — *И доказательств никаких не требуется, — ответил профессор и заговорил негромко, причём его акцент почему-то пропал* (М. Булгаков). Контролируемое целенаправленное усилие в качестве причины *пропасть* невозможно: **От многочисленных фонетических упражнений его акцент наконец пропал*.

Ср. также *Нанесите на жирное пятно неразбавленное средство, и пятно исчезнет за пятнадцать минут*, но не **пропадет за пятнадцать минут*. Поэтому глагол *пропасть* не указывает на прекращение существования таких объектов, которые в силу прагматики не могут восприниматься как исчезающие без причины: **Цивилизация пропала, *Динозавры пропали*. Однако в некоторых случаях *пропасть* допускает указание причины, при условии, что это не чье-то целенаправленное усилие: *После этого случая желание с ним видеться пропало*.

Особенностью глагола *пропасть* по сравнению с *исчезнуть* является употребление применительно к разным сбоям в физиологическом функционировании человека, поскольку они предполагают нарушение естественного хода вещей, нехарактерное для *исчезнуть*; ср. *Сон пропал*, *Аппетит пропал*, *Голос пропал*, *Молоко пропало*, *Месячные пропали*, при странности ?*Сон/аппетит/голос исчез*, ?*Молоко исчезло*, ?*Месячные исчезли*. В подобных контекстах для *пропасть* достаточно характерно указание причины, которой являются различные неконтролируемые воздействия, обычно эмоциональные или физические состояния; ср. примеры из НКРЯ: *От злости у меня пропал дар речи* (А. Сурикова). *Юля часами лежала под капельницей. От переживаний у меня пропало молоко* («Домовой», 2002.01.04).

Корпусные примеры по запросу ***от* S, Gen + *пропасть* (расстояние до 3 слов) + S, Nom (расстояние до двух слов)** в качестве субъекта *пропасть* содержат *голос* (9), *дар речи* (2), *молоко* (1), *слезы* (1), а в качестве причины — *волнение* (6), *ужас* (1), *неожиданность* (1), *злость* (1), *страх* (1), *переживания* (1), *удивление* (1), *слабость* (1).

Улетучиться и *испариться* похожи на *пропасть* в том смысле, что допускают только нецеленаправленные воздействия в качестве причины: ср. **В результате работы с психологом ее страхи улетучились/ испарились*; при возможности *При одном взгляде на его любящее лицо ее страхи улетучились/ испарились*. В ситуациях отсутствия причины выражается нарушение ожиданий говорящего: *Ее веселость куда-то улетучилась/ испарилась*.

Растаять обычно предполагает внешнее воздействие как причину: *От его ласковых слов все ее сомнения сразу растаяли*, но не ?*Ее сомнения почему-то растаяли*.

У глаголов *кончиться* и *иссякнуть* причиной прекращения существования является исчерпанность самого ресурса [Гак 2002:52]; у глагола *иссякнуть* дополнительно выделяется как причина отсутствие ожидаемой подпитки, возобновления ресурса: *Родник иссяк*, но не **кончился*. В ситуации, когда возможны оба глагола, например, *Деньги кончились/иссякли* между ними есть различие: второе сочетание указывает на то, что источник денег перестал их поставлять. Первое сочетание никаких новых денежных вливаний не предполагает: так, можно сказать в магазине *Все, у меня кончились деньги, пошли домой*, но не ?*Все, у меня деньги иссякли*. По признаку причины *кончиться* и *иссякнуть* противопоставлены глаголам *перевестись* и *сойти на нет*, которые часто указывают на то, что нечто, постепенно сокращаясь, как бы само собой превратилось в ничто.

Умереть, *отмереть*, *вымереть*, *вывестись* часто предполагают наличие каких-то причин прекращения существования: либо внутренних причин в организме (*умереть, отмереть*), либо внешних воздействий (*вымереть*, *вывестись*). *Пасть* также обычно предполагает действие внешних факторов.

Во фраземе *кануть в вечность* валентность причины инкопорирована — это воздействие времени и хода истории, заставляющие объект исчезнуть, а окружающих — забыть о нем. *Сгинуть* также часто предполагает в качестве причины прекращения существования объекта ход событий, обычно неблагоприятных, некий водоворот жизни, истории и судьбы: *Бывшие владельцы перемёрли, сгинули кто куда* (Ю. Трифонов); *Так и сгинул он навсегда под мёртвой кличкой: «Номер сто восемнадцатый из первого корпуса»* (М. Булгаков).

Уйти и *пройти* противопоставлены по признаку причины. *Уйти* может указывать как на беспричинное, так и на вызванное внешним воздействием прекращение существования: *Любовь почему-то ушла, и даже дружбы не осталось*; *Выпил таблетку — и боль ушла*, однако обычно не указывает на прекращение существования как результат естественного хода событий: ?*Через неделю простуда сама уходит*. Для *пройти*, напротив, этот тип употреблений, наряду с указанием на внешнее воздействие, очень естественен: *Выпил таблетку — и боль прошла*, *Через неделю простуда сама проходит*. Поэтому для *пройти* естественна сочетаемость с обозначениями болезней, в том числе и метонимическими: *Грипп прошел, Горло прошло, Спина прошла*. Ср. также контрастные фразы *Стеснительность постепенно ушла* vs. *Стеснительность постепенно прошла*, где первая фраза скорее описывает исчезновение свойства, а вторая — исчезновение состояния, рассматриваемое как естественное развитие событий с течением времени.

2.3. Скорость и степень прекращения существования

По этому признаку основные синонимы группы — *исчезнуть* и *пропасть* — также противопоставлены. Эти два синонима предполагают весьма различающиеся структуры события, с чем связано и рассматриваемое ниже различие в фигуре наблюдателя, и их грамматические — в первую очередь аспектуальные — свойства.

Исчезнуть в значении прекращения существования может указывать на постепенный предельный процесс, приводящий к результату (в некоторых других значениях этот глагол устроен совершенно иначе); ср. *Иллюстрирует эту истину и судьба Homo Erectus. Его популяция постепенно исчезла, а ареал ее обитания заселил анатомически современный человек* (А. Волков).

Поэтому глагол НЕСОВ *исчезать* имеет процессные значения, а *исчезать-исчезнуть* представляют собой классическую предельную видовую пару. В обоих видах этот глагол сочетается с адвербиалами типа *постепенно, мало-помалу, один за другим*. При этом *исчезнуть* может указывать и на мгновенное прекращение существования: *Но вот мы въехали во владения графа Модена, и весёлость моя мгновенно исчезла* (А. Н. Апухтин). Таким образом, *исчезнуть* моделирует разные аспекты ситуации прекращения существования: и постепенные процессы, которые к этому приводят, и моментальный результат.

Сам процесс может быть как длительным, так и кратким, а результат как полным, так и частичным; ср. *Продолжительность бывает различна; иногда сыпь исчезает уже через несколько часов, иногда же она исчезает медленно в течение нескольких дней* (Скарлатина и ее гомеопатическое лечение (1911)); *Леопард на Кавказе полностью исчез, полосатых гиен почти не осталось* («Русский репортер», № 15 (143), 22–29 апреля 2010, 2010).

Глагол *пропасть* имеет другую семантическую структуру: он описывает исчезновение как моментальное событие и фокусирует внимание только на результате. Поэтому любая оценка скорости исчезновения невозможна, т. к. процесс, приводящий к результату, остается полностью за кадром; ср. неправильность **быстро/медленно/ постепенно/пропасть*. Соответственно, пара *пропадать-пропасть* устроена иначе: оба глагола описывают результат, у *пропадать* отсутствуют процессные значения; ср. невозможность ??*Пятно постепенно пропадало*, при возможности *Пятно постепенно исчезало.* Для *пропасть* также невозможен частичный результат: ??*Пятно почти/полностью пропало*.

Моментальность *пропасть* семантически связана с беспричинностью, предполагаемой этим глаголом: если есть возможность наблюдать приводящий к чему-либо процесс, результат этого процесса воспринимается как обоснованный. Если процесс остается за кадром, то результат воспринимается как неожиданный и немотивированный.

Прочие глаголы также делятся на те, которые концептуализуют событие исчезновения как постепенное (*сойти на нет*; *кончиться, умереть*; *отмереть*; *вымереть*, *вывестись*, *иссякнуть, перевестись*; *уйти, пройти*; *улетучиться*, *испариться*, *растаять*) и как моментальное (*сгинуть*; *кануть*; *пасть*). Для первых характерно наличие формы НЕСОВ с процессным значением, для вторых — отсутствие формы НЕСОВ.

При этом для некоторых глаголов из первого типа процессность характерна в большей степени, чем для других. Так, у глаголов *сходить на нет, умирать, отмирать, вымирать* форма НЕСОВ имеет естественную процессную интерпретацию, в то время как у *иссякать, переводиться* она более естественно интерпретируется результативно; в частности поэтому для них характерно

употребление с отрицанием; ср. меньшую естественность *Поток желающих приехать постепенно иссякает*, *Богатыри на Руси, к сожалению, переводятся, Тараканы уже выводятся* и естественность *Поток желающих приехать не иссякает*, *Богатыри на Руси не переводятся, Тараканы никак не выводятся*.

Глагол *кончаться* вполне возможен в процессной интерпретации, однако фокусирует внимание лишь на самом последнем, кратком фрагменте прекращения существования: возможно *Хлеб уже кончается*, но менее естественно *Хлеб постепенно кончается, Хлеб медленно кончается*.

Уходить и *проходить* могут иметь как результативную, так и процессную интерпретацию: *Они, случается, ссорятся и бывают несправедливы в споре, и всё же их взаимные обиды уходят без следа* (В. Гроссман); *Да, можно понять, почему говорят, что любовь проходит тогда, когда не о чем говорить* (М. Гиголашвили) vs. *Боль прорезалась. Восторг проходил* (И. Грекова); *Постепенно суета стихала, волнение уходило* (А. Архангельский).

Формы НЕСОВ *улетучиваться*, *испаряться* встречаются в данном значении существенно реже, чем их корреляты совершенного вида, и чаще описывают моментальные события, чем постепенные процессы (в отличие от основного, физического значения); ср. странность *Его страх уже потихоньку улетучивается/испаряется*, при нормальности *Его страх мгновенно улетучивается/испаряется*. Глагол *растаять-таять* вообще редок в данном значении, поэтому затруднительно сделать какие-то обобщения в его отношении.

2.4. Наличие наблюдателя

В отношении наличия наблюдателя синонимы *исчезнуть* и *пропасть* также противопоставлены. Вообще 'прекращение наблюдаемости' — это признак, разлитый по всей полисемии этих двух глаголов, однако неодинаково и неравномерно. В значении 'прекращение существования', как и в других, он отчасти связан с тем, как это прекращение концептуализуется — как процесс или как результат. В целом оба глагола предполагают прекращение существования, влекущее за собой прекращение наблюдаемости.

Однако поскольку *исчезнуть* может описывать очень длительные процессы, которые невозможно охватить в один раунд наблюдения, этот глагол часто указывает на прекращение существования объектов при отсутствии всякого наблюдателя; ср. *Все древние цивилизации исчезли*; *Каждый год исчезает какой-то язык*. В этих примерах речь идет о таких объектах, прекращение существования которых в силу тех или иных причин наблюдать невозможно. В этом смысле *исчезнуть* может описывать чистое прекращение существования, без компонента визуального наблюдения.

Для глагола *пропасть*, который фокусирует внимание на результате, характерно употребление в контекстах, описывающих прекращение существования видимых объектов. При этом фразы типа *Прыщи пропали, Веснушки пропали*; *Красные пятна на лице начали пропадать* значат, что говорящий некоторое время назад наблюдал *прыщи, веснушки и пятна*, а в момент речи видит

их отсутствие, т. е. наблюдает лишь часть события — результат, — но не приводящий к нему процесс.

Подобные контексты характерны и для *исчезнуть*; *исчезают прыщи, морщины, бородавки* и пр. Однако в *исчезнуть* наблюдатель может видеть как процесс, так и результат; в этом смысле ему близок глагол *сходить* (*Пятна постепенно сходят, Пятна полностью сошли*). Однако *исчезнуть* позволяет наблюдать процесс как бы под еще большим увеличением, чем *сходить*; ср. *Снег исчезал на глазах*, *Пятно исчезло прямо на глазах*, при меньшей естественности *Снег сходил на глазах, Пятно сошло прямо на глазах*.

Оба глагола также могут описывать внутренние, т. е. часто не заметные наблюдателю, состояния, которые тем не менее имеют перцептивный компонент, т. к. воспринимаются экспериенцером изнутри: Страх *исчез*, *Желание исчезло* vs. *Страх пропал, Желание пропало*. Однако здесь симметрии между ними нет: для *исчезнуть* такое употребление гораздо более характерно: *Мечты/надежды/подозрения/вопросы исчезли*; *Тревога/тоска исчезла*; *Раздражение исчезло*, при странности ?*Мечты/надежды/ подозрения/вопросы пропали*. Впрочем, это различие связано с признаком причины и ожидания говорящего.

Для *пропасть* характерен более узкий круг контекстов, описывающих неожиданное и часто неприятное для говорящего прекращение состояния: *Интерес к учебе у детей пропал*; *Желание с ним общаться полностью пропало*. В этом употреблении между ними также сохраняется различие в структуре события: экспериенцер *исчезнуть* может осознавать постепенное исчезновение какого-либо состояния, в то время как экспериенцер *пропасть* фиксирует лишь внезапно наступивший и поэтому неожиданный для него результат.

Что касается остальных глаголов, то наличие или отсутствие наблюдателя в них определяется, во-первых, типом объекта, во-вторых, длительностью или моментальностью события. Глаголы, описывающие постепенное, но не слишком длительное прекращение существования материальных объектов, могут предполагать наблюдателя: *Он умер у меня на глазах*; *Смотри, яблоки кончились*. Все прочие глаголы обозначают либо исчезновение нематериальных объектов, либо моментальные, либо слишком длительные события, т. е. не предполагают наблюдаемости перехода от существованию к несуществованию.

Литература

1. *Апресян 2001* — Ю. Д. Апресян. Системообразующие смыслы ‘знать’ и ‘считать’ в русском языке // Русский язык в научном освещении. 2001. № 1. с. 5–26.
2. *Апресян 2006* — Ю. Д. Апресян. Фундаментальная классификация предикатов // Языковая картина мира и системная лексикография / Под ред. Ю. Д. Апресяна. М., 2006. с. 75–109.
3. *Гак 2002* — В. Г. Гак. Семантическое поле *конца*. // Логический анализ языка. Семантика начала и конца. Отв. ред. Н. Д. Арутюнова. М., 2002. с. 50–55.

4. *Кобозева 1993* — И. М. Кобозева. Мысль и идея на фоне категоризации ментальных имен. // Логический анализ языка. Ментальные действия. М., 1993. с. 95–104.
5. *Кустова 2002* — Г. И. Кустова. Семантические аспекты лексических функций (глаголы со значением 'начаться'/'кончиться') // Логический анализ языка. Семантика начала и конца. Отв. ред. Н. Д. Арутюнова. М., 2002. с. 69–82.
6. *Падучева 2004* — Е. В. Падучева. Динамические модели в семантике лексики. М., 2004.

References

1. *Apresjan 2001* — Apresjan Ju. D. System-forming meanings 'to know' and 'to think' in Russian [Sistemoobrazujushchie smysly 'znat' i 'schitat' s russkom jazyke]. Russian language in linguistic perspective. [Russkij jazyk v nauchnom osveshchenii]. 2001. N 1. pp. 5–26.
2. *Apresjan 2006* — Apresjan Ju. D. Fundamental classification of predicates [Fundamental'naja klassifikacija predikatov]. Linguistic worldview and systematic lexicography [Jazykovaja kartina mira i sistemnaja leskikografija]. Ju. D. Apresjan, ed. Moscow, 2006. pp. 75–109.
3. *Gak 2002* — Gak V. G. Semantic field of 'end' [Semanticheskoe pole *kontsa*]. Logical analysis of language. Semantics of beginning and end. [Logicheskii analiz iazyka. Semantika nachala i kontsa]. N. D. Arutjunova, ed. Moscow, 2002.
4. *Kobozeva 1993* — Kobozeva I. M. Thought and idea against the backdrop of categorizing mental nouns [Mysl' i ideja na fone kategorizatsii mental'nyx imion]. Logical analysis of language. Mental actions. [Logicheskii analiz iazyka. Mental'nye dejstvija]. N. D. Arutjunova, N. K. Rjabtseva, eds. Moscow, 1993. Pp. 95–104.
5. *Kustova 2002* — Kustova G. I. Semantic aspects of lexical functions (verbs meaning 'to begin'/'to end') [Semanticheskie aspekty leksicheskix funktsij (glagoly so znacheniem 'nachat'sja'/ 'konchit'sja')]. Logical analysis of language. Semantics of beginning and end. [Logicheskii analiz iazyka. Semantika nachala i kontsa]. N. D. Arutjunova, ed. Moscow, 2002. Pp. 69–82.
6. *Paducheva 2004* — Paducheva E. V. Dynamic models in the semantics of the lexicon. [Dinamicheskie modeli v semantike leksiki]. Moscow, 2004.

СЕМАНТИКА И ПРАГМАТИКА *ПОСЛЕДНЕГО* И *ПРЕДПОСЛЕДНЕГО*[1]

Апресян В. Ю. (valentina.apresjan@gmail.com)
Национальный исследовательский университет Высшая школа экономики; Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Шмелев А. Д. (shmelev.alexei@gmail.com)
Московский педагогический государственный университет; Институт русского языка им. В. В. Виноградова РАН; Православный Свято-Тихоновский гуманитарный университет, Москва, Россия

В работе рассматривается система значений прилагательного *последний*. Предлагается семантический инвариант прилагательного *последний* с двумя семантическими валентностями — элемента и последовательности — и демонстрируется, каким образом модификации инварианта, включая добавление валентностей ориентира и момента времени, приводят к образованию новых значений. Показывается, каким образом семантические свойства *последнего* в каждом из его значений мотивируют его синтаксические и сочетаемостные свойства. Кроме того, обсуждается роль прагматики и лексикализации грамматических форм и конструкций в разрешении неоднозначности сочетаний со словом *последний*, а также лексические соответствия этого прилагательного в английском языке.

Ключевые слова: семантика, прагматика, валентность, конструкция, сочетаемость, статические последовательности, динамические последовательности, временной отрезок, прагматическая импликатура

[1] Работа над данной статьей отчасти финансировалась следующими грантами: грант Программы фундаментальных исследований Президиума РАН «Историческая память и российская идентичность» — «Основной лексический фонд русского языка как элемент русской культуры: системная организация лексики и ее отражение в словаре» (2015–2017), грант РГНФ № 16-04-00302 «Подготовка третьего выпуска Активного словаря русского языка» (2016–2018), грант РГНФ, проект № 15-04-00488 «Изменения узуса и кодификация норм русского литературного языка» (2015–2017).

SEMANTICS AND PRAGMATICS OF THE RUSSIAN WORDS *POSLEDNIJ* AND *PREDPOSLEDNIJ*

Apresjan V. Ju. (valentina.apresjan@gmail.com)
National Research University Higher School of Economics; Vinogradov Russian Language Institute of the Russian Academyof Sciences, Moscow, Russia

Shmelev A. D. (shmelev.alexei@gmail.com)
Moscow Pedagogical State University; Vinogradov Institute of Russian Language, Russian Academy of Sciences; St Tikhon’s Orthodox University, Moscow, Russia

The paper considers the senses of the Russian adjective *poslednij* ‘last’. Its polysemy is analyzed as deriving from a certain core semantic structure that is common to all its meanings. The core structure has two semantic valencies — of a sequence and of a sequence element. Modifications of the core structure, including additional valencies (point of reference and landmark) account for its polysemy, as well as for diversity of its collocational and syntactic properties. The paper also demonstrates the role of pragmatics and lexicalization of grammatical and syntactic forms in disambiguating different meanings of *poslednij*, against the backdrop of its English correlates.

Key words: semantics, pragmatics, valency, construction, collocation, static sequence, dynamic sequence, time period, pragmatic implicature

1. Семантический инвариант *последнего*

В работе предлагается трактовка русского прилагательного *последний*, позволяющая описать разные его значения как единую систему, а также объяснить его синтаксические и сочетаемостные свойства в разных значениях. Некоторые из рассматриваемых в статье языковых фактов обсуждаются в работе [Спиридонова 2002], однако в целом их представление в данной работе существенным образом отличается. В первую очередь это касается предлагаемой структуры полисемии и семантического инварианта *последнего*, а также толкований значений и анализа механизмов семантической деривации. У *последний* есть семантический инвариант, который повторяется во всех значениях. Его можно сформулировать следующим образом:

Последний X в Y = ‘Элемент X расположен в последовательности однородных элементов Y таким образом, что все остальные элементы Y идут до него’.

Как видно из этого толкования, у *последний* есть две валентности — элемента и последовательности; ср. *последнее слово* [X] *в строке* [Y]. Интерпретация *последнего* зависит от характера последовательности и ее элементов.

Почти для всех сочетаний с *последним* возможна интерпретация ‘такой X, после которого уже не будет другого такого Xa’; ср. *Этот разговор оказался последним*.

В этой интерпретации добавляется валентность предела Z, в рамках которого не будет повторения Xa, а валентность последовательности, как правило, не реализуется: *На сегодня* [Z] *это последняя конфета* [X], *Сегодня последний семинар* [X] *в этом году* [Z]. Если валентность предела не выражена, в данной интерпретации под пределом по умолчанию понимается конец жизни (об интерпретации абсолютный конец’ см. [Спиридонова 2002:174]).

Для интерпретации ‘последний в жизни’ характерна рематичность слова *последний*; ср. акцентное выделение: *Это наша* ***последняя*** *встреча*.

Прагматическая неоднозначность сочетаний типа *последняя встреча* и наличие лакуны в русском языке, не позволяющее разрешить неоднозначность лексически[2], объясняют «суеверное» использование слова *крайний* с целью разрешения неоднозначности; ср. просторечные высказывания типа *крайний прыжок с парашютом* = ‘последний на данный момент’.

В этой интерпретации *предпоследний*[3] является возможным, хотя и не частым аналогом *последнего*, поскольку предполагает знание того, что произойдет или произошло: *Илья Иосифович, давно уже освобождённый из* ***предпоследней****, как выяснилось впоследствии, отсидки* […]*, разглагольствовал о тайном шифре жизни …* (Л. Улицкая).

2. Типы последовательностей и значения *последнего*

Последовательности могут быть статическими, т. е. законченными (*строка, таблица, алфавит*) и динамическими, т. е. продолжающимися (*новости, семинары, выпуски газеты*). Один тип последовательности может переходить в другой — так, последовательность *книг* писателя является динамической в течение его жизни и становится статической после его смерти. Ниже рассматриваются разные употребления слова *последний*, модификации семантического инварианта его значения и синтаксических способов выражения валентности в зависимости от типа последовательности и создающих ее объектов.

---

[2] Ср. английское *last meeting* (‘в жизни или в каком-то ее отрезке’) vs. *latest meeting* (‘на данный момент’).

[3] Это же касается и более редкого аналога *предпредпоследний* и всех потенциальных дериватов, образованных по этой модели.

2.1. Статические последовательности

Статические последовательности — это такие, в которых число элементов не изменяется. Они делятся на фасадные и не фасадные [о семантическом признаке фасадности см. Fillmore 1969, Апресян 1974:111–112, Апресян 1995:40–42]. Под фасадными последовательностями понимаются такие, у которых порядок расположения элементов фиксирован, т. е. начало и конец последовательности не зависят от положения наблюдателя и других объектов в пространстве.

2.1.1. Фасадные статические последовательности

Прототипические фасадные последовательности — это текстовые объекты: *абзацы, строчки, таблицы*. Для них в русскоязычной культуре фиксирован порядок слева направо и сверху вниз, в соответствии с правилами чтения и письма. Таким образом, *последняя строчка в абзаце*– это самая нижняя строчка, *последнее слово на строчке* — самое правое.

Для *книг* принцип упорядочивания несколько другой. Скажем, *последняя страница книги* — эта та, которая имеет наибольший порядковый номер (при условии, если они все пронумерованы) или самая нижняя при естественном расположении книги в ее использовании, т. е. когда книга находится в горизонтальном положении на некоторой поверхности и ее передняя обложка (та, на которой напечатано название) обращена вверх. Эта страница останется *последней*, даже если книгу перевернуть или если начать чтение *с последней страницы*.

Помимо текстовых объектов, фасадны также некоторые пространственные последовательности. Например, *сидеть на последней парте* однозначно интерпретируется как ‘сидеть на той парте, которая расположена дальше всего от доски’. Так же устроены сочетания типа *сидеть на последнем ряду в театре* (на самом удаленном от сцены).

Сочетание *на последнем этаже* тоже понимается однозначно — как ‘на самом высоком этаже дома’, потому что в русскоязычной культуре этажи нумеруются, начиная от самого близкого к земле.

Что касается сочетаний типа *последний дом*, то их интерпретация неоднозначна: если имеется в виду нумерация, то речь идет о фасадной статической последовательности, и *последний дом на улице* — это имеющий наибольший номер; если говорящий воспринимает дома визуально — это наиболее удаленный от него дом. Кроме того, мысленно располагая дома относительно наблюдателя, находящегося в середине улицы, можно говорить о *последнем доме слева* и *последнем доме справа*. Во всех этих случаях речь идет о нефасадной последовательности. Таким образом, очевидно, что деление на фасадные и нефасадные последовательности до некоторой степени условно, т. к некоторые фасадные последовательности могут интерпретироваться нефасадно.

При употреблении *последнего* в контексте фасадных статических последовательностей валентность последовательности Y может выражаться зависимым при существительном, выражающем валентность X: *последний этаж (дома), последний этап (конкурса)*.

В подобных контекстах может употребляться и слово *предпоследний*. По данным ресурса Sketch Engine (корпус RuTenTen), именно статические последовательности формируют самый частый тип употреблений *предпоследнего*: *предпоследний слог, абзац, строчка, строка, столбец, этаж, этап, раунд*. Это весьма естественно: в статической последовательности *предпоследний*, как и любое другое закрепленное пронумерованное место в ряду, легко фиксируется.

2.1.2. Нефасадные статические последовательности

Нефасадные последовательности упорядочиваются в зависимости от положения наблюдателя или ориентира. Нефасадные последовательности могут создаваться любыми объектами, не предполагающими фиксированного порядка; чаще всего — это стоящие в ряд пространственные объекты или люди. *Последний* в подобных контекстах часто используется для объяснения, о каком из объектов идет речь. Для устранения неоднозначности вводится указание на ориентир (им может быть наблюдатель) и/или направление упорядочивания объектов: *Но я заметил, что одно кресло, самое последнее справа от меня, пустует* (В. Бережков); *Домик монтажниц торчал четвертым слева на последней от штаба улице* (В. Корнилов).

Употребление *последнего* в подобных контекстах ограничено наличием у него сильного конкурента — слова *крайний*; ср. *Казарин стоит крайним справа* (Д. Быков). Так, в НКРЯ встретилось 8 сочетаний *последний справа* и 28 сочетаний *крайний справа*.

Впрочем, *крайний* применим только к горизонтально расположенным последовательностям; если последовательность расположена вертикально, употребляется *последний*: *последняя ступенька*, но не **крайняя ступенька*. Сочетание *последняя ступенька* представляет собой интересный контраст сочетанию *последний этаж*: если *последний этаж* — это всегда самый верхний, интерпретация *ступеньки* зависит от направления движения наблюдателя; ср. *Добрался до последней ступеньки и упёрся лбом в потолок* [самая верхняя] (Ю. Домбровский); *Он спустился с последней ступеньки и зашёл в кабинет напротив* [самая нижняя] (Ю. Домбровский).

Валентность последовательности в этом типе употреблений *последнего* часто не выражается, поскольку сама последовательность может не иметь специального названия; ср. избыточность *последняя девушка (в ряду людей)*, *последний дом (в ряду домов)*. При этом добавляется обязательная валентность ориентира и/или направления отсчета объектов: *последний справа от центра/ от леса, последний сверху/снизу/ справа/слева*.

В подобных контекстах также употребляется *предпоследний*; ср. *предпоследняя ступенька, Он на фотографии предпоследний справа*.

2.1.3. Создающиеся статические последовательности

Промежуточный случай между статическими и динамическими последовательностями образует тип **создающихся** статических последовательностей. В каждый заданный момент времени они представляют собой статическую фасадную последовательность, однако во времени число и порядок элементов

могут меняться. Соответственно, статус *последнего* у объектов, образующих эти последовательности, является временным. Создающиеся статические последовательности — это очереди, а также порядок следования в разного рода соревнованиях; ср. *последний в очереди; последний в рейтинге*.

У данного типа употреблений валентность элемента X обычно не выражена, валентность Y выражена предложно-именной группой при *последнем*: *последний (спортсмен) в турнирной таблице*, *последний (человек) в очереди*. Само слово *последний* обычно употребляется копредикативно в творительном падеже, т. е. в депиктивной конструкции, что подчеркивает временный и ситуативный статус *последнего* (о депиктивной конструкции см. [Кузнецова, Рахилина 2010]): *выступать последним, последним, прийти последним, стоять последним, закончить последним*.

Сдвинутый тип употреблений в этом значении образуют сочетания *последнего* с кместоимением и сослагательным наклонением вида *Он последний, к кому бы я обратилась за помощью*, где очередь выстраивается лишь в сознании говорящего.

В этих контекстах также может употребляться *предпоследний*: *стоять предпоследним*, *идти предпоследним в турнирной таблице*. Однако в конструкции с кместоимением слово *предпоследний* не употребляется; ср. невозможность **Он предпоследний, к кому бы я обратилась за помощью*.

2.2. Динамические последовательности

Динамические последовательности делятся на такие, в которых объекты исчезают, и такие, в которых объекты добавляются.

2.2.1. Динамические последовательности с исчезающими объектами

Типичные последовательности с исчезающими объектами — это разного рода ресурсы, которые постепенно исчерпываются, так что остается только один объект или часть — *последний* или *последняя*; ср. *Он съел последнее яблоко*; *Он другу последнюю рубашку отдаст*; *сражаться из последних сил/ до последней капли крови*; *последние соки/ крохи* (такие сочетания часто склонны к идиоматизации). К этому же типу употреблений относятся *последний шанс, последняя возможность*, которые также воспринимаются как ограниченный ресурс.

Валентность последовательности в этом значении обычно не выражается, поскольку названием динамической последовательности является название ее объектов во множественном числе; ср. невозможность тавтологического высказывания **Он съел последнее яблоко из яблок*. Однако существует возможность выразить ее при помощи выделительной конструкции с предлогом *из*, если не эксплицировать валентность элемента; правда, это возможно не всегда и требует добавления прилагательного: *Он съел последнее из оставшихся яблок*, но не **Он съел последний из оставшегося сахара*. Чтобы употребление этой конструкции было возможно, необходимо, чтобы элементы динамической последовательности были дискретными и составляли некоторое множество.

В этом значении *предпоследний* также допустим (*Он съел предпоследний банан*), но только в ситуациях, когда элементы дискретны и подвергаются реальному счету; соответственно, сочетания с названиями совокупностей невозможны: **предпоследний сахар, *предпоследние бананы*. Также невозможна замена *последний* на *предпоследний* в идиоматизированных сочетаниях, где речь не идет о реальном последовательном исчезновении объектов: **Он отдаст другу предпоследнюю рубашку*.

При этом в совокупности исчезающих объектов некоторый объект можно назвать *последним* еще до того, как он исчез (*На тарелке осталось последнее яблоко*), а *предпоследним* объект оказывается только после своего исчезновения (из двух яблок, лежащих на тарелке, ни одно не является *предпоследним*).

2.2.2. Динамические последовательности с появляющимися объектами

Этот класс представлен регулярно появляющимися информационными объектами; ср. — *выпуски газет, музыкальные альбомы, фильмы, книги, версии программ* и пр. Инвариант значения *последнего* модифицируется в этой интерпретации как: ‘на момент речи после данного объекта других подобных объектов не было’.

Отдельная разновидность контекстов внутри этой интерпретации — это последовательности со сменяющими друг друга объектами, где последовательность образуется разными вариантами объекта, причем каждый последующий отменяет остальные; ср. *последнее завещание* — такой вариант завещания, после которого на момент речи новые изменения не вносились и который является на момент речи актуальным. Ср. также *последний вариант статьи*, *последнее решение правительства*, *последняя редакция книги*. Синонимом *последнего* в этом значении является *актуальный*, антонимами — *предыдущий, предпоследний, предшествующий*, английский коррелят — *latest*.

Если известно, что новые варианты объекта по той или иной причине появляться не будут, динамическая последовательность превращается в статическую: *Это мое последнее решение* [не на данный момент, а вообще — оно больше не будет меняться]. В этой интерпретации синоним *последнего* — *окончательный*, английский коррелят — *final*. К этому же типу употреблений относятся сочетания *последняя истина, последняя прямота*, *последняя правота*.

Выбор интерпретации определяется прагматикой; ср. *Я ей писал, что не оставлю ее у отца, и она знает, что это мое* ***последнее*** ‘окончательное’ *решение* (А. К. Шеллер-Михайлов) vs. *Таким образом,* ***последнее*** ‘актуальное на данный момент’ *решение Багдада снижает вероятность того, что ОПЕК на этой неделе примет решение об увеличении квот на добычу нефти* («Финансовая Россия», 2002.09.19).

Валентность последовательности в этом значении обычно не выражается, за исключением выделительной конструкции с предлогом *из*: *последний из выпусков «Троицкого варианта»*. Кроме того, возникает валентность момента времени, которая часто не выражается эксплицитно; ср. *последняя (на тот момент) книга Улицкой*. В дейктическом режиме, если эта валентность не выражена, высказывание интерпретируется относительно момента речи.

Слово *предпоследний* употребляется в динамической интерпретации ‘актуальный’, но не в статической ‘окончательный’; ср. *На её вопрос: «Ведь моё **последнее** завещание действительно, а **предпоследнее** нет?» — её нотариус отвечал: «Валентина Михайловна, в вашем случае я бы ставил на документах час»* (С. Спивакова), но нельзя: ?*Это мое предпоследнее решение*, **Я хочу знать предпоследнюю истину*.

2.2.3. Динамические последовательности с событиями и мероприятиями

Типичные последовательности в этом круге употреблений –регулярные события и мероприятия: *последнее землетрясение, последний семинар, последний разговор, последняя встреча, последний матч* (об этом круге употреблений см. [Спиридонова 2002: 172]). Для этого круга употреблений важна повторяемость и ожидаемость событий; ср. естественные сочетания *последнее наводнение/ торнадо/ землетрясение/солнечное затмение*, *последний снегопад/дождь* (о погодных явлениях, которые предсказывают метеослужбы), но странно ??*последний иней,* ??*последняя роса*.

Как отмечалось выше, все подобные сочетания могут интерпретироваться статически — ‘такое событие или мероприятие, которое больше не повторится в рамках заданного отрезка времени или никогда’ (*Эта встреча оказалась последней*). Что касается динамических интерпретаций, то здесь возможны два варианта.

Рассмотрим фразу *На нашем последнем семинаре мы обсуждали максимы Грайса*. Эту фразу можно сказать как во время происходящего в данный момент семинара, и тогда *последний* будет значить ‘предыдущий’ и будет являться синонимом *предпоследнего, предыдущего* и антонимом *нынешнего, сегодняшнего* (инвариант значения *последнего* модифицируется в этой интерпретации как: ‘после того события и до события сейчас не было другого подобного события’). Однако она может быть сказана и до того, как следующий семинар состоится, например, по телефону или в письме, тогда это ‘такое событие, после которого на момент речи еще не было другого такого события’.

Отдельно в рамках данного употребления интересно рассмотреть сочетание *(в) последний раз* [о некоторых различиях предложных и беспредложных синтаксических фразем со словом *раз* и прилагательным см. Иомдин 2016]. Хотя вообще русский язык лексически не различает значения ‘на данный момент ситуация больше не повторялась’ и ‘ситуация больше никогда не повторится’, на отдельных сочетаниях та или иная интерпретация может быть лексикализована. Примером такой лексикализации являются сочетания *последний раз* и *в последний раз*: первое обычно задает интерпретацию ‘на данный момент ситуация больше не повторялась’, второе — ‘ситуация больше никогда не повторится’ (это не правило, поскольку для каждого сочетания возможны противоположные интерпретации, а тенденция). В силу этого различия *последний раз* тяготеет к употреблению в теме, а *в последний раз* — в реме: *Последний раз мы виделись в прошлую пятницу* vs. *В прошлую пятницу мы виделись в последний раз*. С отрицанием для обоих вариантов возможна только интерпретация ‘ситуация больше никогда не повторится’: *Не последний раз/Не в последний раз видимся*.

Валентность последовательности в этом значении обычно не выражается, возможна выделительная конструкция: *последний из матчей сезона*. Как у всех динамических последовательностей, есть факультативная валентность момента времени. Если она не выражена, то высказывание интерпретируется относительно момента речи. *Предпоследний* употребляется в подобных контекстах: *Предпоследний матч они проиграли*.

2.2.4. Динамические последовательности с отрезками времени

Едва ли не самыми характерными для прилагательного *последний* оказываются сочетания с обозначениями отрезков времени. В соответствии с подразделением, описанным в статье [Шмелев 2011][4], разграничиваются три класса обозначений отрезков времени, которые условно названы «длительностями», «фиксированными периодами» и «циклами». «Длительности» дают отрезку времени чисто количественную характеристику: они характеризуют его лишь с точки зрения его продолжительности, безотносительно к его положению на временной оси. К ним относятся такие слова, как *минута 1*, *час 1*, *день 1*, *сутки 1*, *неделя 1*, *месяц 1*, *год 1*. «Фиксированные периоды» характеризуют отрезки времени с точки зрения их положения на временной оси, это такие слова, как *понедельник*, *15 сентября*, *февраль*, *зима*, *утро*, *день 2*. Большинство «фиксированных периодов» циклически повторяется, т. е. представляют собою части «циклов». К «циклам» относятся *день 3* (продолжающийся от полуночи до полуночи), *месяц 2*, длящийся с 1го по 28е, 29е, 30е или 31е число, *неделя 2*, состоящая из *понедельника*, *вторника* и т. д., *год 2*, начинающийся 1 января и заканчивающийся 31 декабря. Главная особенность «цикла» заключается в том, что, как только цикл заканчивается, начинается новый цикл.

Обратим внимание на то, что во многих случаях обозначения «длительностей» и обозначения «циклов» внешне совпадают. Это следует иметь в виду при обсуждении интерпретации таких сочетаний, как *последний год*, *последний месяц*, *последняя неделя*.

Забегая вперед, заметим, что многие обозначения отрезков времени обладают идиосинкратическими свойствами в отношении интерпретации их сочетаний со словом *последний*. В частности, в сочетании со многими существительными прилагательное *последний* получает различную интерпретацию в зависимости от того, употреблено ли существительное в форме единственного или множественного числа.Опишем некоторые правила интерпретации сочетаний прилагательного *последний* с обозначениями временных отрезков.

Для всех сочетаний, включающих обозначение «длительности», возможна интерпретация **‘последняя часть целого’**, т. е. отрезок указанной длительности, непосредственно предшествующий окончанию этого целого: *последний*

---

[4] В указанной статье данная классификация использовалась для описания сочетаемости обозначений отрезков времени с операторами, отсылающими к будущему, т. е. с прилагательными *будущий*, *следующий*, *ближайший*, *предстоящий*. Однако очевидно, что она имеет значительно более общий характер и релевантна для описания языкового поведения рассматриваемых разных единиц, сочетающихся с обозначениями отрезков времени.

год тюремного заключения, последняя неделя беременности, последние минуты матча. При этом окончание целого к моменту речи может быть как состоявшимся, так и запланированным. Разновидностью данной интерпретации является понимание прилагательного как **‘последний в жизни’**: *последний год Толстого*. Данная разновидность в норме возможна только после смерти того, о ком идет речь; тем самым характеристика *последний* дается post factum[5]. Сюда же примыкает интерпретация **‘непосредственно предшествующий какому-либо событию’**: *последний перед эмиграцией год, последний день перед отъездом, последняя минута перед расставанием*, когда в фокусе внимания оказывается не период, частью которого является рассматриваемый отрезок, а событие (как правило, начинающее некоторый новый период). Особый случай представляет собою дейктическая интерпретация, когда имеется в виду отрезок, непосредственно предшествующий моменту речи или иной дейктической точке отсчета: *Последний месяц я плохо себя чувствую* (примерно 30 дней перед моментом речи).

Для «фиксированных периодов» существуют сходные интерпретации: ‘часть целого, после которой в пределах этого целого не было аналогичного фиксированного периода’ (*последний четверг января*); ‘последний в жизни’ (*Та Пасха оказалась последней в жизни Романовых*)[6]; ‘последний перед каким-либо событием’ (*последняя суббота перед выборами*); дейктическая интерпретация (*Вот в последнее воскресенье укладываю ее, а она просит посидеть с ней еще…* [Михаил Шишкин]).

Для циклов сочетания со словом *последний* не очень характерны. Это связано с тем, что обозначения циклов внешне совпадают с обозначениями длительностей и в сочетании со словом *последний* обычно интерпретируются как длительности. Для указания на цикл предпочтительными оказываются сочетания со словом *прошлый* (если цикл уже завершился) или *прошедший*, *истекший* (они возможны и в рамках незавершенного цикла).

Лексикализация форм числа различается для разных временных отрезков. Так, для слова *день* дейктическая интерпретация оказывается вероятной для формы множественного числа (*Последние дни я плохо себя чувствую*), но маловероятной для формы единственного числа (??*Последний день я плохо себя чувствую*)[7], а для слова *час* дело обстоит противоположным образом (*Последний час я сижу и скучаю*, но едва ли ??*Последние часы я сижу и скучаю*).

---

[5] Впрочем, это определяется внелингвистическими соображениями: смерть редко планируется заранее. Однако можно найти контексты, в которых интерпретация окончания жизни как запланированного события оказывается вполне возможной или даже предпочтительной. Предложение *Луиза Талбот решила провести последний день своей жизни, сидя в тени большого сикомора у заборчика перед ее домом* (вольный перевод начального предложения детективного романа, цитированного в статье [Шмелев 1997: 472] для иллюстрации противопоставления de dicto vs. de re при передаче чужих мнений или высказываний) наводит на мысль о запланированной смерти. Лишь продолжение, сообщающее, что Луиза не знала, что ей предстоит умереть, снимает это предположение.

[6] Эта разновидность для календарных периодов не характерна.

[7] Это связано с тем, что для смысла ‘день накануне момента речи’ используется слово *вчера*. Точно так же и смысл, соответствующий английскому *last evening* используется выражение *вчера вечером*.

Уместно упомянуть также слово *время*. В единственном числе для него более всего характерна дейктическая интерпретация ‘неопределенное время, непосредственно предшествующее моменту речи или иной точке отсчета’ (*Последнее время я плохо себя чувствую*), а во множественном числе слово обычно понимается как ‘время, непосредственно предшествующее концу света’ (*Настали последние времена*).

Слово *предпоследний* в рассматриваемых контекстах используется весьма ограниченно; в частности, для него не характерно дейктическое употребление.

3. Оценочные употребления *последнего*. Сочетаемостные особенности, семантическая мотивация

Существует устойчивая культурная конвенция упорядочивания объектов от «важных» к «неважным» и от «лучших» к «худшим». С этой конвенцией связаны возникающие у высказываний со словом *последний* прагматические импликатуры ‘плохой’ и ‘неважный’. Конвенционализация этих импликатур приводит к появлению у слова *последний* значений ‘самый незначительный’ и ‘очень плохой’.

Первое иллюстрируется сочетаниями *не из последних удальцов, не последний специалист в своей области*. Оно характеризуется отрицательной поляризацией, так как реализуется только в сочетании с частицей *не* и чаще всего с обозначениями лиц, обладающих положительными свойствами или компетенциями.

Второе значение иллюстрируется сочетаниями *последнее дело, ругаться последними словами, последний негодяй, последняя сволочь, последний дурак*. При сочетании с нейтральными словами (*дело, слова*) *последний* в этом значении задает отрицательную оценку для всего сочетания; при сочетании со словами, уже включающими отрицательную оценку (*негодяй, сволочь, дурак*), она усиливается.

В следующем примере из «Правой кисти» Александра Солженицына мы имеем дело с диффузным употреблением, когда трудно однозначно определить, какое из двух значений имеется в виду: *Последняя из этих женщин не решилась бы пройтись со мною рядом!* Слово *предпоследний* не имеет аналогичных значений.

Литература

1. *Апресян* 1974 — Ю. Д. Апресян. Лексическая семантика. Синонимические средства языка. М., 1974.
2. *Апресян* 1995 — Ю. Д. Апресян. Интегральное описание языка и системная лексикография. М., 1995.
3. *Иомдин* 2016 — Л. Л. Иомдин. Конструкции микросинтаксиса, образованные русской лексемой раз (в печати).
4. *Кузнецова,* Рахилина 2010 — Ю. Л. Кузнецова, Е. В. Рахилина. Русские депиктивы // Рахилина Е. В. (отв. ред.). Лингвистика конструкций. М., 2010. С. 159–183.

5. *Спиридонова* 2002 — Н. Ф. Спиридонова. От начала к концу: семантика прилагательного последний. // Арутюнова Н. Д. (отв. ред.). Логический анализ языка. Семантика начала и конца. М., 2002. С. 169–180.
6. *Шмелев* 1997 — А. Д. Шмелев. Приемы языковой демагогии. Апелляция к реальности как демагогический прием // Булыгина Т. В., Шмелев А. Д. Языковая концептуализация мира (на материале русской грамматики). М., 1997. С. 461–477.
7. *Шмелев* 2011 — А. Д. Шмелев. Парадоксы референции к будущему // Арутюнова Н. Д. (отв. ред.). Лингвофутуризм. Взгляд языка в будущее. М., 2011. С. 288–301.
8. *Fillmore* 1969 — Ch. J. Fillmore. Types of lexical information // Kiefer F. (ed.). Studies in syntax and semantics. Dordrecht, 1969. pp. 109–137.

References

1. *Apresjan Ju. D.* (1974), Lexical semantics. Synonymic Means of Language [Leksicheskaia semantika. Sinonimicheskie sredstva iazyka], Nauka, Moscow.
2. *Apresjan Ju. D.* (1995), Integral description of language and systemic lexicography. [Integral’noe opisanie iazyka i sistemnaia leksikografiia], Shkola Iazyki Russkoi Kul’tury, Moscow.
3. *Iomdin L. L.* (in print), Microsyntactic constructions formed by the Russian word *raz* [Konstruktsii mikrosintaksisa, obrazovannye russkoj leksemoj *raz*].
4. *Kuznetsova Ju. L., Rakhilina E. V.* (2010), Russian depictives [Russkie depiktivy]. Rakhilina E. V. (ed.) Linguistics of constructions. [Lingvistika konstruktsij], Azbukovnik, Moscow, pp. 159–183.
5. *Spiridonova N. F.* (2002), From the beginning to the end: the semantics of the adjective *poslednij* ‘last’ [Ot nachala k kontsu: semantika prilagatel’nogo *poslednij*], Arutjunova N. D. (ed.) Logical analysis of language. Semantics of beginning and end. [Logicheskii analiz iazyka. Semantika nachala i kontsa], Indrik, Moscow, pp. 169–180.
6. *Shmelev A. D.* (1997), Methods of linguistic manipulation: Reference to “reality” as a manipulative practice [Priemy iazykovoi demagogii: Apelliatsiia k real’nosti kak demagogicheskii priem], Bulygina T. V., Shmelev A. D. Linguistic conceptualization of the world (from the evidence of the Russian grammar) [Iazykovaia kontseptualizatsiia mira (na materiale russkoi grammatiki)], Shkola Iazyki Russkoi Kul’tury, Moscow, pp. 461–477.
7. *Shmelev A. D.* (2011), Paradoxes of reference to the future [Paradoksy referentsii k budushchemu], Linguistic futurism: Linguistic view of the future [Lingvofuturizm: Vzgliad iazyka v budushchee], Indrik, Moscow, pp. 288–301.
8. *Fillmore Ch. J.* (1969), Types of lexical information, Kiefer F. (ed.). Studies in syntax and semantics, Reidel, Dordrecht, 1969, pp. 109–137.

DEVELOPING A POLYSYNTHETIC LANGUAGE CORPUS: PROBLEMS AND SOLUTIONS[1]

Arkhangelskiy T. A. (tarkhangelskiy@hse.ru),
Lander Yu. A. (yulander@hse.ru)

National Research University Higher School of Economics, Moscow, Russia

Although there exist comprehensive morphologically annotated corpora for many morphologically rich languages, there have been no such corpora for any polysynthetic language so far. Developing a corpus of a polysynthetic language poses a range of theoretical and practical challenges for corpus linguistics. Some of these challenges have been partly addressed when developing corpora for languages with extensive morphological inventories and numerous productive derivation models such as Turkic or Uralic, while others are unique for this kind of languages. As we are currently working on a corpus of the polysynthetic West Circassian language, we had to identify these challenges and propose theoretical and practical solutions. These include the tokenization problem, which involves delimiting morphology from syntax, the problem with lemmatization and part-of-speech tagging, and a number of glossing and search issues. The solutions proposed in the paper are partly implemented and will be available for public testing when the preliminary version of the corpus is released.

Keywords: polysynthesis, Adyghe, West Circassian, language corpora, morphology

РАЗРАБОТКА КОРПУСА ПОЛИСИНТЕТИЧЕСКОГО ЯЗЫКА: ПРОБЛЕМЫ И РЕШЕНИЯ

Архангельский Т. А. (tarkhangelskiy@hse.ru),
Ландер Ю. А. (yulander@hse.ru)

Национальный исследовательский университет «Высшая школа экономики», Москва, Россия

Несмотря на то, что в настоящее время существует множество морфологически размеченных корпусов для языков с богатой морфологией, до сих пор не было создано ни одного корпуса полисинтетического языка,

[1] This paper is based on our research supported by Russian Foundation for Basic Research (grant No. 15-06-07434a).

который бы учитывал необходимую морфологическую информацию. Разработка корпуса для таких языков ставит перед корпусным лингвистом ряд нетривиальных теоретических и практических задач. Некоторые из них в меньшем объёме встречались и частично решались ранее при создании корпусов языков с большими морфологическими системами и обилием продуктивных деривационных моделей, например, тюркских или уральских языков. Однако многие из этих проблем уникальны для полисинтетических языков. В ходе работы над созданием корпуса полисинтетического адыгейского языка мы обрисовываем эти проблемы и предлагаем ряд теоретических и практических решений. Описываемые проблемы включают в себя токенизацию (связанную с нечёткой границей между синтаксисом и мофологией), лемматизацию и морфологическую разметку, а также ряд вопросов, связанных с глоссированием и поиском в корпусе. Предлагаемые решения частично реализованы и будут доступны для тестирования в пилотной версии корпуса.

Ключевые слова: полисинтетизм, адыгейский язык, языковые корпуса, морфология

1. Introduction

The main feature that distinguishes a corpus of language from a mere collection of texts is its annotation. While it is possible to add various levels of annotation to a corpus, including syntactic parsing, semantic labeling, anaphora resolution, etc., what is absolutely necessary for morphologically rich languages is morphological annotation. Traditionally, this kind of annotation is split into lemmatization and tagging. Lemmatization means annotating words with their lemmata (dictionary forms). The term tagging generally means annotating words with grammatical tags, such as “noun” or “genitive case”. While tagsets of moderately morphologically complex languages often only include part-of-speech tags, tagsets for more complex ones usually cover all morphological categories. The annotation is normally searchable with the help of an online or offline search interface.

Since corpora of written languages are normally too large for manual annotation to be feasible, their compilation includes developing an automatic tool for morphological annotation. Naturally, the more complex the morphology, the more difficult it is to develop such a tool. However, that when the complexity reaches certain point, we come face to face with completely new challenges beyond the increase in size of the formalized description of the morphology. It turns out that the very concepts of lemmatization and tagging have to be redefined to embrace the complexity of the morphological system. While these problems appeared to a certain degree and were partially addressed in corpora of such languages as Turkish, Tatar or Udmurt, they primarily manifest themselves in polysynthetic languages. In this paper, we outline these challenges, using the data from West Circassian (also known as Adyghe), a polysynthetic language belonging to the Circassian branch of the Northwest Caucasian (Abkhaz-Adyghe) family. To the best of our knowledge, no publicly available morphologically annotated corpora of polysynthetic languages have been developed to date, which makes our corpus unique.

Polysynthetic languages can be informally described as languages that may convey morphologically much of the information that in standard synthetic languages like Russian is conveyed by syntax. Consider the following West Circassian example:

(1)	*...тыгъужъыми*	*ыцэхэр*	*къыфыIуигъэпсыгъэх*
	təʁʷəʔə-m-jə	ə-ce-xe-r	qə-fə-ʔʷ-jə-ʁe-psə-ʁe-x
	wolf-OBL-ADD	3SG.PR-tooth-PL-ABS	DIR-BEN-LOC-3SG.ERG-CAUS-shine-PST-PL
	‘and the wolf	made its teeth	shine for him’

The verb in (1) simultaneously contains not only the causative and cross-reference affixes but also a locative preverb and the benefactive applicative, which here introduces the null cross-reference affix of the beneficiary.

Not surprisingly, such languages differ from Standard Average European in many respects. For example, much of their morphology is highly productive and shows syntactic properties (e.g., recursion, semantically based variation in order, etc.); cf. de Reuse (2009) who coined the term “productive non-inflectional concatenation” (PNC) for this kind of morphology. In addition, polysynthetic morphology sometimes is not at all selective and can attach to stems belonging to various lexical classes. For example, in West Circassian, tense markers appear not only on verbs, but also on adjectives, nouns and even postpositions. These properties of morphology pose multiple problems for tagging polysynthetic texts, as will be shown below.

2. West Circassian corpus

Developing West Circassian corpus is an ongoing project that started in 2015. As West Circassian is a written language with standard orthography based on the Cyrillic alphabet, the corpus will mostly consist of written texts, however, a certain number of manually annotated spoken texts collected during fieldwork will also be added. The corpus is being built in line with the general principles of medium-scale corpus design developed within the framework of Russian Academy of Sciences Corpus Program which was in effect in 2011–2014[2]. The workflow adapted in this program includes collecting written texts in standard orthography, developing an automated morphological analysis tool, annotating the texts with it and placing the corpus in the search engine with publicly available search interface. Morphological tagging in this framework is usually rule-based, being carried out with the help of a formalized description of the inflection and productive derivation of a language together with a grammatical dictionary containing the description of its lexis. The search engine which was used for most of these corpora and which we are going to use in the West Circassian corpus, was originally developed for the Eastern Armenian National Corpus and by default allows search by wordform, lemma or stem, translation, a combination of grammatical tags, as well as complex search involving a combination of the aforementioned properties (see Arkhangelskiy

[2] Most middle-scale corpora developed within the framework of this program are available at http://web-corpora.net.

et al. 2012). However, the search capabilities which were sufficient for non-polysynthetic languages, proved to be insufficient for the West Circassian data, and have to be enhanced. We replaced the "bag of tags" principle, according to which the morphological tagger assigns each token grammatical tags without specifying relative order of tags within the set, with a mechanism that allows specifying relative position of morphemes in a search query. This enhancement is discussed in detail in section 3.4.

Currently, we are testing the solutions proposed below with a pilot version of grammatical dictionary. A publicly available preliminary version of the corpus is expected to be released in 2016.

3. Problems and solutions

For any West Circassian token, the following types of morphological and lexical annotation are included in our corpus:

(i) lemmatization,
(ii) part-of-speech attribution,
(iii) the presence of productive morphemes,
(iv) the order of productive morphemes.

Here productive morphemes comprise both inflection and PNC but not non-regular derivation which should be covered by the lexicon. The discussion will cover these topics in that order.

3.1. Tokenization: the subtle boundary between syntax and morphology

Tokenization, which is the first task in the text processing pipeline, already poses a problem specific for West Circassian. There admittedly exist difficulties for tokenization even in non-polysynthetic languages, e.g. annotation of multiword named entities (such as "New York"), contractions, hyphenated words or text-based emoticons, as well as ways for dealing with these difficulties (cf. Grana et al. 2002, Bocharov et al. 2012). Most existing corpora assume for the sake of technical simplicity that a token cannot contain a whitespace, thus disregarding named entities (or annotating them at a separate level) and offering solutions for other problems within the limits of this constraint. Indeed, splitting the text into pieces delimited by whitespaces before further processing makes the tokenization step relatively fast and easy.

Although West Circassian normally does distinguish between syntactic relations and relations between morphemes, there are certain problems in demarcating morphology and syntax which lead to another kind of tokenization difficulties. Consider the following Adyghe example (2):

(2) *иджэнэ шхъонтIэ дахэхэр*
jə-ʒene-šχʷenṭe-daxe-xe-r
POSS-dress-blue-beautiful-PL-ABS
'her beautiful blue dresses'

This example consists of three graphical tokens separated by whitespaces in standard orthography. Although it looks like an ordinary noun phrase, phonetic and morphologic criteria (specifically the absence of e/a alternation in the nominal stem, see Arkadiev, Testelets 2009) indicate that this *nominal complex* behaves as a single wordform (see Lander, to appear (a) for details). The reason why this is so problematic for corpus construction is the following. When attached to a nominal complex, prefixes and suffixes normally go to the left and the right edges of the whole complex, respectively. For instance, in the example (2) above, the plural marker modifies the whole complex apparently headed by the noun 'dress'. However, if only graphical tokens are taken into account when performing morphological analysis, search queries like "dress in plural" or "a combination of a possessive and a plural marker in one word" will miss this example.

The nominal complex problem has no simple solution. If we do not recur to machine learning or other statistical methods which require a manually tagged golden standard corpus, all rule-based methods will not provide accurate results, as most words do not have alternations, and in most cases not having any prefixes or suffixes is perfectly normal for a West Circassian word. Even if we can identify such complexes accurately enough, annotating the whole complex as a single token has its drawbacks. For example, a simple query like "the token *daxexer*" would not find this graphical token inside a complex. At the current stage, we are not including nominal complexes recognition in our tokenization module. However, in the process of morphological analysis we are tagging tokens with no expected alternation, which can help in recognizing complexes in the future.

3.2. Lemmatization

The idea that a lemma can be attributed to every or almost every word is usually taken for granted in contemporary corpus linguistics. While this statement undoubtedly holds for all major languages for which corpora have been created, the situation is much less clear for languages with productive derivational morphology. Turkic or some of the Uralic languages provide "light" versions of such challenge which have been addressed in corpora and in bilingual dictionaries. In these languages, multiple derivational affixes, specifically, verbal markers such as causatives or iteratives, or nominalizations, may attach to the stem. Although these affixes are very productive and mostly semantically regular, in some cases they add to the meaning of the word in a non-compositional way. One of the existing solutions to this problem is annotating roots instead of lemmata. Another option is providing two alternative variants of tagging, which allows users to search for both derived and non-derived lemmata. Although this leads to some morphological ambiguity, its scale is limited in these languages: for example, according to Khakimov et al. (2014), this kind of ambiguity accounts for only 7.2% of all ambiguously tagged tokens in Tatar National Corpus.

In polysynthetic languages, however, this problem is much more pervasive and profound. In West Circassian, there are plenty of PNC affixes which are so productive that it is infeasible to include any new item derived with them into the lexicon. Nevertheless, the derived items often have non-compositional meanings, with the meanings themselves being often far less predictable than in Turkic languages.

Consider, for example, the applicative derivation, which adds an indirect object to the subcategorization frame of a word (see Smeets 1984; Lander, to appear (b); Lander & Letuchiy, to appear for details). West Circassian possesses a dozen of applicative affixes which may be added to roots and stems in a straightforward manner, as in (3) where the benefactive complex translates in English as 'for them':

(3) *афэтшIыщт*
[a-fe]-t-ṣ̂ə-š't
3PL.IO-BEN-1PL.ERG-do-FUT
'We will do this for them.'

Since the applicative *fe-* is highly productive, its semantic contribution is purely compositional here, and it can easily be omitted (resulting in the form *tṣ̂əš't* 'we will do this'), it makes no sense to lemmatize the form with this prefix.

The situation is different in (4), though.

(4) *фэмышIыгъэ*
fe-mə-ṣ̂ə-ʁe
BEN-NEG-do-PST
'not prosperous'

In this negative form of the word *fe-ṣ̂ə-ʁe* BEN-do-PST 'prosperous', only the negative prefix is used compositionally. The contribution of the benefactive applicative prefix and the past tense suffix is, on the other hand, idiomatic, despite the fact that both affixes are fully productive and are usually not likely to construct new lexemes. In languages like West Circassian, this kind of idiomatic lexicalization of morpheme combinations is quite widespread. It is evident that this situation requires consistent treatment that would go beyond the ambiguous analysis solution discussed above. Apart from search-related concerns, treating such combinations as having multiple ambiguous analyses with different stems or lemmata leads to difficulties during morphological tagging, as in West Circassian combinations of the root and derivational affixes can be split by inflectional morphemes. This would necessarily require adding disjointed stems to the dictionary and dealing with non-concatenative morphology, which makes morphological tagging a much more difficult task, although not completely impossible.

To address this problem, we use two different levels of annotation which are filled one after the other. During the main stage of tagging, the tokens are split into morphemes and glossed. Thus, every successfully analyzed token is assigned a lemma coinciding with its root, the description of which is stored at the first level. Then, the annotated token is passed to the second-level annotation module. This module loads a YAML file with rules that look like "if a token has root X together with affixes X, Y and Z, it should be assigned secondary lemma L". After applying the rules to the first level of annotation, all possible lemmata are written to the second level. For the word in the example (4), the first level will contain only information about the primary lemma *ṣ̂ən* 'do'. At the second level, it will be also associated with the lemma *feṣ̂əʁe* 'prosperous'. The search interface, correspondingly, was adapted to perform queries on both primary and secondary lemma at the same time.

3.3. Parts-of-speech (POS) tagging

The same kind of problems we face in lemmatization leads to challenges for POS-tagging as well. As with lemmatization, these challenges are present in Turkic and Uralic languages, to a much lesser extent. Specifically, these languages often have productive nominalization suffixes which can be used to derive a noun from virtually any verbal stem. Within the ambiguous analyses framework described above, the problem can be solved by assigning different POS tags to different analyses: the analysis that has the bare stem as its lemma will be assigned the tag "Verb", and the one where lemma includes the nominalization affix, the tag "Noun". Another way of addressing this issue, offered, for example, by Sak et al. (2008) for Turkish, is treating POS tags just like ordinary morpheme tags. In this approach, the stem and every POS-changing morpheme is annotated with the corresponding POS tag and, consequently, the analysis of one token can have more than one POS tag.

The situation is much more difficult in polysynthetic languages. Because of low selectivity of many affixes, the word class distinction itself is a serious problem for such languages[3]. In West Circassian, for example, tense affixes may attach to clearly nominal stems. The question is, then, whether this tense marker derives a new verb (see Lander and Testelets 2006 for some evidence) or it is simply not associated with any specific POS. Since both decisions are not theoretically fully justified in this case, we prefer to abstain from attempting to determine the POS tag of the word as a whole and rather only specify the POS of its primary lemma.

Note that many wordforms with derivational affixes still are likely to be analyzed as belonging to one of the parts of speech, due to the presence of affixes that may be considered as clearly defining the class of the derived item. Examples of such affixes include the causative prefix and the agentive nominalization illustrated in (5):

(5)	*уагъэшIущт*	*къекIокIакIо*
	w-a-ʁe-ŝʷə-šʼt	q-je-ḳʷe-čʼ-aḳʷe
	2PL.ABS-3PL.ERG-CAUS-good-FUT	DIR-DAT-go-go.out-AG
	'they will humour you (lit. make you good)'	'vagrant'

Nevertheless, even in the presence of such morphemes it is not always possible to unambiguously assign one of the POS tags to the token. For instance, when both the causative prefix and the nominalization suffix are present, it is not clear what applies the first and what applies the second. For example, in (6a) the causative clearly applies to the nominalization, but in (6b) the nominalization applies to the causative, as shown by brackets:

(6) a. *Зыжъугъэбэнакlу!*
zə-ẑʷ-ʁe-[ben-akʷ]
RFL.ABS-2PL.ERG-CAUS-[fight-AG]
'Make yourselves fighters!'

---

[3] For different views on the issue see, for example, Baker 2004 and several papers in Rijkhoff and van Lier (eds.) 2013. For Circassian see Lander and Testelets 2006.

b. *А гъэрэхьэтакIор сэры!*
a [ʁe-šx]-aḳʷe-r se-rə
that [CAUS-console]-AG-ABS I-PRED
'That consoler is me!'

In order to enable searching for tokens for which it is possible to define a single POS tag, we suggest tagging affixes which clearly indicate the part of speech with additional labels such as NOMINAL or VERBAL. Such tagging will allow searching for e. g. all tokens which can be safely analyzed as nominal, by automatically transforming the query into "find all tokens which have a stem or a derivational affix marked as nominal and no derivational affixes marked as verbal". At the same time, the decision will make it possible to look for any roots with any derivational suffixes, without specifying the final, resulting POS attribution.

3.4. Glossing and search capabilities

In nearly all large automatically tagged corpora each token is annotated with what is called 'a bag of tags', without specifying number of occurrences of each tag or their relative order. This approach is fully justified for Standard Average European languages, however it is hardly appropriate for West Circassian. One of the obstacles is recursion, whereby one affix or group of affixes may be used more than once during derivation (cf. Lander and Letuchiy 2010), as in example (7) below, which contains two benefactive applicative prefixes. The first of them introduces the argument corresponding to the phrase 'for them' in the English translation, and the second introduces the recipient argument translated as 'to him/her':

(7) *сафыфэтхэ*
s-a-fə-Ø-f-e-txe
1SG.ABS-3PL.IO-BEN-3SG.IO-BEN-DYN-write
'I write to him/her for them'

Another obstacle is variable morpheme order. While in some cases order may be irrelevant, in some others it affects the meaning because of the morpheme scope hierarchies. Finally, it is often important whether two morphemes border each other: e. g. when searching for a combination of an indirect object personal affix with an applicative, like in (7).

In order to address these problems, full glossing rather than a mere set of tags is stored for each token in the database of the corpus engine. We use abstract glosses, which are commonly used by typologists (see Lehmann 1982; Haspelmath 2002: 34–36) and present in the examples above. The query interface, still allowing for the 'bag-of-tag'-style queries, has been enhanced with a "glossing" search field which works with a glossing query language. When designing such a language, we considered the tradeoff between expressive power of the query language and the speed, as overly complex queries are usually hard to implement efficiently.

The query language we propose allows for any number of elementary queries joined by Boolean operators. Each elementary query can include grammatical tags and wildcard characters ? and *, the former standing for exactly one morpheme and the latter standing for any number of morphemes. Left and right word boundaries are marked by #. Morpheme adjacency is indicated by hyphens and their order is taken into account. For example, the query "#DIR-*-PST-?-ADV" will find all words starting with the directive prefix, following any number of morphemes, then a past tense marker, then another morpheme, and then and adverbial case marker. The language also allows using umbrella tags that unite several grammatical tags, e. g. the tag "APP" matches any applicative derivation affix such as BEN in (3), (4) and (7). Such elementary queries are currently transformed to SQL-queries containing regular expressions.

4. Conclusion

Corpus linguists dealing with polysynthetic language data face new kinds of challenges which are characteristic and often unique for these languages. It turns out that for such languages, many traditional techniques and concepts are not directly applicable to the data, and novel ways of text processing and corpus design should be developed.

We identified some of the problems which arise in the course of development of West Circassian language corpus, and offered possible solutions for them. The most important challenges, from our point of view, include somewhat vague boundary between morphology and syntax (hence tokenization problems), not well defined concepts of a lemma and a part of speech, and annotating the texts in such a way to enable search queries that could take into account phenomena like recursion and relative order of morphemes.

It should be noted that in this paper we only focused on a limited number of issues raised during the elaboration of the corpora. Some others include:

(i) morphophonological rules, which are by no means numerous but still should be accounted because of their high relevance for the analysis of the West Circassian word,

(ii) classes of morphemes: as we noted in Section 3.4, there is a need to group morphemes into classes, but the criteria of such grouping remain obscure,

(iii) the "translation" of our system into the conceptual system which is traditionally used in the descriptions of Circassian languages and in textbooks and hence should be considered for practical reasons.

References

1. *Arkadiev, P., Testelets, Ya.* (2009), On three alternations in the Adyghe language [O trex čeredovanijax v adygejskom jazyke], in Ya. G. Testelets et al. (eds), Aspekty polisintetizma: očerki po grammatike adygejskogo jazyka, 121–145. Moscow: RGGU.

2. *Arkhangelskiy, T., Belyaev, O., Vydrin, A.* (2012), The creation of large-scaled annotated corpora of minority languages using UniParser and the EANC platform, in Proceedings of COLING 2012: Posters, 83–91. Mumbai: The COLING 2012 Organizing Committee.
3. *Baker, M. C.* (2004), Lexical Categories: Verbs, Nouns, and Adjectives. Oxford: Oxford University Press.
4. *Bocharov, V., Granovsky D., Surikov, A.* (2012), Probabilistic tokenization model in the Open Corpus project [Verojatnostnaja model' tokenizacii v proékte Otkrytyj korpus], in Novye informacionnye texnologii v avtomatizirovannyx sistemax: Materialy 15, 15, 176–183.
5. *Grana, J., Barcala, F. M., Vilares, J.* (2002), Formal methods of tokenization for part-of-speech tagging, in Computational linguistics and intelligent text processing, 240–249. Springer Berlin—Heidelberg.
6. *Haspelmath, M.* (2002), Understanding Morphology. London: Arnold.
7. *Khakimov, B., Gil'mullin, R., Gataullin, R.* (2014), Morphological disambiguation in the Tatar corpus [Razrešenie grammatičeskoj mnogoznačnosti v korpuse tatarskogo jazyka], in Učenye zapiski Kazanskogo gosuniversiteta 156(5): 236–244.
8. *Lander, Yu.* (to appear, a), Nominal complex in West Circassian: between morphology and syntax, in Studies in Language.
9. *Lander, Yu.* (to appear, b), Adyghe, in P. O. Müller et al. (eds), Word Formation, An International Handbook of the Languages of Europe. Berlin: Mouton de Gruyter.
10. *Lander, Yu., Letuchiy, A.* (2010), Kinds of recursion in Adyghe morphology, in: H. van der Hulst (ed.), Recursion and Human Language, 263–284. Berlin: Mouton de Gruyter.
11. *Lander, Yu., Letuchiy, A.* (to appear), Decreasing valency-changing operations in a valency-increasing language? In Í. Navarro and A. Alvarez (eds), On verb valency change: theoretical and typological perspectives (working title).
12. *Lander, Yu., Testelets, Ya.* (2006), Nouniness and specificity: Circassian and Wakashan. Paper presented at the conference on Universality and Particularity in Parts-of-Speech Systems, University of Amsterdam.
13. *Lehmann, Chr.* (1982), Directions for interlinear morphemic translations, in Folia Linguistica 16: 193–224.
14. *de Reuse, W. J.* (2009), Polysynthesis as a typological feature. An attempt at a characterization from Eskimo and Athabascan perspectives, in M.-A. Mahieu and N. Tersis (eds), Variations on Polysynthesis: the Eskaleut Languages, 19–34. Amsterdam: John Benjamins.
15. *Rijkhoff, J., van Lier, E.* (eds). (2013), Flexible Word Classes, in Typological Studies of Underspecified Parts of Speech. Oxford: Oxford University Press.
16. *Sak, H., Güngör, T., Saraçlar, M.* (2008), Turkish language resources: Morphological parser, morphological disambiguator and web corpus, in Advances in natural language processing, 417–427. Springer Berlin—Heidelberg.
17. *Smeets, R.* (1984), Studies in West Circassian Phonology and Morphology. Leiden: The Hakuchi Press.

СРАВНЕНИЕ АРХИТЕКТУР НЕЙРОННЫХ СЕТЕЙ В ЗАДАЧЕ АНАЛИЗА ТОНАЛЬНОСТИ РУССКОЯЗЫЧНЫХ ТВИТОВ

Архипенко К. (arkhipenko@ispras.ru)[1,2],
Козлов И. (kozlov-ilya@ispras.ru)[1,3],
Трофимович Ю. (integral@ispras.ru)[1],
Скорняков К. (kirill.skorniakov@ispras.ru)[1,3],
Гомзин А. (gomzin@ispras.ru)[1,2],
Турдаков Д. (turdakov@ispras.ru)[1,2,4]

[1]Институт Системного Программирования РАН, Москва, Россия

[2]Московский государственный университет им. М. В. Ломоносова, ВМК, Москва, Россия

[3]Московский физико-технический институт, Долгопрудный, Россия

[4]НИУ Высшая школа экономики, ФКН, Москва, Россия

Ключевые слова: анализ тональности, извлечение мнений, рекурентная нейронная сеть, свёрточная нейронная сеть

COMPARISON OF NEURAL NETWORK ARCHITECTURES FOR SENTIMENT ANALYSIS OF RUSSIAN TWEETS

Arkhipenko K. (arkhipenko@ispras.ru)[1,2],
Kozlov I. (kozlov-ilya@ispras.ru)[1,3],
Trofimovich J. (integral@ispras.ru)[1],
Skorniakov K. (kirill.skorniakov@ispras.ru)[1,3],
Gomzin A. (gomzin@ispras.ru)[1,2],
Turdakov D. (turdakov@ispras.ru)[1,2,4]

[1]Institute for System Programming of RAS, Moscow, Russia

[2]Lomonosov Moscow State University, CMC faculty, Moscow, Russia

[3]MIPT, Dolgoprudny, Russia

[4]FCS NRU HSE, Moscow, Russia

The paper presents evaluation of three neural network based approaches to Twitter sentiment analysis task performed at SentiRuEval-2016. The task focuses on sentiment classification of tweets about banks and telecommunication companies.

Our team submitted three solutions which are based on different supervised classifiers: Gated Recurrent Unit neural network (GRU), convolutional neural network (CNN), and SVM classifier with domain adaptation combined with previous two classifiers. We used vector representations of words obtained with word2vec model as features for classifiers. These classifiers were trained on labeled data provided by organizers of the evaluation. Additionally, we collected several million posts and comments from social networks for training word2vec model.

According to evaluation results, GRU-based solution shows the best macro-averaged F1-score for both domains (banks and telecommunication companies) and also has the best micro-averaged F1-score for banks domain among all solutions submitted to SentiRuEval.

Key words: sentiment analysis, opinion mining, recurrent neural network, convolutional neural network

Introduction

The paper describes participation in SentiRuEval-2016 competition. The task of the competition focuses on object-oriented sentiment analysis of Russian messages posted by Twitter users. The messages are about banks and telecommunication companies.

The goal of the task is detection of sentiment (negative, neutral or positive) with respect to organizations (banks or telecommunication companies) mentioned in Twitter message. Thus it can be viewed as three-class classification task. The organizers of the evaluation provided labeled training datasets along with unlabeled test datasets for both banks and telecommunication companies. Training datasets contain about 9,000 Twitter messages each, while test datasets contain about 19,000 messages each.

In this paper, we focus on detection of overall sentiment of messages. Object-oriented sentiment classification with algorithms used in this paper is a part of our further research.

All variants of our sentiment analysis system use supervised machine learning algorithms. One of our main goals is evaluation of artificial neural networks (ANNs) for sentiment analysis task. In this paper, we evaluate algorithms based on recurrent neural network (RNN) and convolutional neural network (CNN) along with shallow machine learning approach—SVM with domain adaptation. In each of these three cases we use word2vec (Mikolov et al., 2013a) vectors as features for the algorithms.

We have submitted three solutions to SentiRuEval-2016. The first two are based on recurrent neural network and convolutional neural network, respectively. The last solution is an ensemble solution consisting of three classifiers. It uses SVM with domain adaptation along with RNN and CNN.

The paper is organized as follows: Section 1 provides overview of the related work; Section 2 presents full description of our methods and features that we used;

Section 3 provides evaluation results for different methods; in the final section we make conclusion for this work.

1. Related work

Artificial neural networks have become very popular in recent years. They have been shown to achieve state-of-the-art results in various NLP tasks, outperforming shallow machine learning algorithms like support vector machines (SVMs), hidden Markov models and conditional random fields (CRFs).

Recurrent neural networks (RNNs) are considered to be one of the most powerful models for sequence modeling. The success of RNNs in the area of sentence classification was reported by many researchers (Irsoy & Cardie, 2014) (Adamson & Turan, 2015) (Tang et al., 2015).

Convolutional neural networks (CNNs) are another class of neural networks initially designed for image processing. However, CNNs have been shown in recent years to perform very well in NLP tasks, including sentiment analysis and sentence modeling tasks (Kalchbrenner et al., 2014) (Kim, 2014) (dos Santos et al., 2014).

It has been shown that neural network based models for NLP become especially powerful when they are pre-trained with some vector space model (Collobert et al., 2011). The most common way to do this is to use distributed representations of words. The most popular such model now is word2vec (Mikolov et al., 2013a), which improves many NLP tasks.

2. Method description

2.1. Word2vec

Word2vec (Mikolov et al., 2013a) (Mikolov et al., 2013b) is a popular model for computationally efficient learning vector representations of words. Vectors learned using word2vec have been shown to capture semantic information between words (Mikolov et al., 2013c), and pre-training using word2vec leads to major improvements in many NLP tasks.

We used original word2vec toolkit[1] for obtaining vector representations of Russian words. The model was trained on 3.3 GB of user-submitted posts from VK, LiveJournal, echo.msk.ru and svpressa.ru. All the text was lowercased, and punctuation was removed. The following parameters were used for learning:

1. Continuous Bag-of-Words (CBOW) architecture with negative sampling (10 negative samples for every prediction);
2. vector size of 200;
3. maximum context window size of 5;

[1] https://code.google.com/archive/p/word2vec/

4. 5 training iterations over corpus;
5. words occurring in the corpus less than 25 times were discarded from the vocabulary; the resulting vocabulary size was 249,014.

2.2. Recurrent neural network

Recurrent neural networks (RNNs) are a class of neural networks that have recurrent connections between units. This makes RNNs well-suited to classify and predict sequence data, including short documents.

Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) is a popular RNN architecture designed to cope with long-term dependency problem. LSTM has been shown to achieve state-of-the-art or comparable to state-of-the-art results in many text sequence processing tasks (Sutskever et al., 2014) (Palangi et al., 2015).

Gated Recurrent Unit (GRU) (Cho et al., 2014) is a simplified version of LSTM that has been shown to outperform LSTM in some tasks (Chung et al., 2014), although according to (Jozefowicz et al., 2015) the gap between LSTM and GRU can often be closed by changing initialization of LSTM cells.

Our RNN-based model is composed of LSTM/GRU network regularized by dropout with probability of 0.5 and succeeded by fully connected layer with 3 neurons that predict probabilities of each class—negative, neutral and positive. The input sample is lowercased and converted to sequence of corresponding word2vec vectors described in section 2.1. Punctuation and words that are not in word2vec vocabulary are discarded. The resulting sequence of vectors is input to the network. Like word2vec vectors, the size of input and output of LSTM/GRU cells is 200.

We tried several variations of recurrent networks: shallow LSTM/GRU, bidirectional GRU and two-layer GRU. We also tried to revert the order of input vector sequences.

We used Keras library[2] to implement the model[3]. In case of LSTM initialization of cells recommended in (Jozefowicz et al., 2015) was used. Sigmoid and hard sigmoid were used for recurrent network as output activation and hidden activation, respectively; softmax was used as activation of fully connected layer.

Adam optimizer (Kingma & Ba, 2014) and batch size of 8 were used for training; the number of training epochs was set to 20.

2.3. Convolutional neural network

Due to widely reported success of CNNs (convolutional neural networks) (Kalchbrenner et al., 2014) (Kim, 2014) (dos Santos et al., 2014) in the area of sentiment analysis we have conducted some experiments with CNN as well.

---

[2] https://github.com/fchollet/keras

[3] Source code is available on https://github.com/arkhipenko-ispras/SentiRuEval-2016-RNN

We used word2vec word vectors described in section 2.1 as features. For each tweet the matrix S is constructed where s_i (i-th row) is a word vector for the i-th word in tweet. Then we calculate two vectors t^{avg} and t^{max} as follows:

$$t_j^{avg} = \frac{1}{m} \sum_{1 \le i \le m} s_{ij} \tag{1}$$

$$t_j^{max} = \max_{1 \le i \le m} s_{ij} \tag{2}$$

Concatenation of these two vectors is input to our CNN. The network is composed of convolutional layer with 8 kernels of width 10 which is succeeded by dense layer with 3 neurons (with softmax activation) that predict probabilities of each class. scikit-neuralnetwork library[4] was used for implementing the network. The number of training epochs was set to 10.

The roadmap for further survey includes experiments not only with different kinds of features but also with architecture of the CNN as well. Feature extraction with word2vec seems to be the most promising one. Since CNNs are not as powerful in sequence processing as RNNs the technique of Dynamic k-Max Pooling (Kalchbrenner et al., 2014) can be used to address the problem of variable sentence length.

2.4. Domain adaptation and ensemble solution

2.4.1. Domain adaptation

In most cases we assume that source domain (train data) and target domain (test data) are driven from the same probability distribution:

$$P_s(X, y) \equiv P_t(X, y) \tag{3}$$

Consequently this means that it is impossible to build classifier that would be able to distinguish target domain sample from source domain sample. But in many real world problems assumption (3) does not hold and

$$P_s(X, y) \neq P_t(X, y) \tag{4}$$

How one can detect that $P_s(X, y) \neq P_t(X, y)$?

1. Quality of the model, measured on source domain (e.g. with cross-validation) is much higher than on the target domain. Some participants of SentiRuEval-2015 faced this problem.
2. Consequence of assumption (3) is impossibility to build classifier which can distinguish target domain from source domain. The ability to build

[4] https://github.com/aigamedev/scikit-neuralnetwork

such classifier indicates that assumption (3) does not hold. We were able to achieve F1-score on source vs target domain classification above 0.85.

One can improve quality of algorithm in target domain with different method of domain adaptation. Some methods can be found in (Jiang, 2008).

In this work we use a simple method of domain adaptation—sample reweighting. Let $l(x, y, \theta)$ be a loss function. In order to obtain θ we want to minimize following function:

$$L(\theta) = \sum_{x,y \in X \times Y} (x, y, \theta) P_t(x, y) \to \min_{\theta} \tag{5}$$

We can write function L in the equivalent form:

$$L(\theta) = \sum_{x,y \in X \times Y} (x, y, \theta) \frac{P_t(x, y)}{P_s(x, y)} P_s(x, y) \tag{6}$$

Now replace true loss function with empirical estimation:

$$\hat{L}(\theta) = \frac{1}{l} \sum_{i=1}^{l} (x_i, y_i, \theta) \frac{P_t(x_i, y_i)}{P_s(x_i, y_i)} \tag{7}$$

As one can see that algorithm leads as to the feature reweighting with $w_i = \frac{P_t(x_i, y_i)}{P_s(x_i, y_i)}$. Finally we assume that $P_t(y|x) \equiv P_s(y|x)$, thus weight w_i can be found as $w_i = \frac{P(x_i|t)}{P(x_i|s)}$. With Bayes' theorem one can estimate weight as:

$$w_i = \frac{P(t|x_i)P(s)}{P(s|x_i)P(t)} = C \times \frac{P(t|x_i)}{P(s|x_i)} \tag{8}$$

We estimate weight with the logistic regression, and it slightly increases the quality.

2.4.2. Our ensemble solution

Our ensemble classifier consists of three classifiers; each of them votes with equal weight. The first two are GRU neural network and convolutional neural network described in sections 2.2 and 2.3, respectively.

The third classifier is SVM with sample reweighting described in 2.4.1. We used polynomial kernel with degree of 3. For every tweet, the average of word2vec vectors (described in section 2.1) of all words in the tweet is used as features for the SVM classifier.

3. Evaluation

Tables 1–2 present results of the evalution on sentiment classification. Both tables show macro-averaged F1-score of negative and positive classes, used as quality measure on SentiRuEval-2016 competition.

For recurrent neural network based model, we performed 5-fold cross-validation on training data provided by organizers of SentiRuEval. The results are showed in Table 1. We found that GRU network slightly outperforms LSTM network, and that reversing the order of words in tweets improves the quality. Adding an extra recurrent layer also slightly increases the quality.

In addition, we found that using word2vec vectors as features for recurrent network is crucial. Using randomly initialized embedding layer and one-hot features instead of word2vec features gives macro-averaged F1-score of only 0.45 for banks and 0.47 for telecommunication companies.

Table 1. Macro-averaged F1-score, evaluated with RNN models using 5-fold cross-validation on SentiRuEval training data

RNN Architecture	Domain	
	Banks	Telecommunication companies
LSTM	0.6026	0.6410
GRU	0.6129	0.6428
GRU, reversed sequences	0.6211	0.6570
Bidirectional GRU	0.6207	0.6521
Two-layer GRU, reversed sequences	0.6243	0.6597

Table 2. F1-score and ranks among all solutions, evaluated on SentiRuEval test data (according to SentiRuEval results)

Classifier	Domain			
	Banks		Telecommunication companies	
	Macro (score/rank)	Micro (score/rank)	Macro (score/rank)	Micro (score/rank)
CNN	0.4832 / 21	0.5253 / 21	0.4704 / 41	0.6060 / 36
Two-layer GRU, reversed sequences	0.5517 / 1	0.5881 / 1	0.5594 / 1	0.6569 / 21
Ensemble classifier	0.5352 / 2	0.5749 / 2	0.5403 / 9	0.6525 / 23
Best solution not from our team	0.5252 / 3	0.5653 / 3	0.5493 / 2	0.6822 / 1

Table 2 shows results on SentiRuEval test datasets for solutions described in sections 2.2–2.4. It also shows micro-averaged version of F1-score and includes solutions' ranks among all 58 solutions submitted to SentiRuEval by 10 teams. For test data classification with GRU network, the model was trained on whole train data 5 times and correspondingly gave 5 predictions for test data. Then the leading class over all predictions was chosen for each sample. Other models were trained and predicted once.

The Gated Recurrent Unit based solution got the best macro-averaged score on both domains, significantly outperforming solutions from other teams on banks domain, and also has the best micro-averaged F1-score on banks domain.

Conclusion

We have described all variants of our sentiment analysis system. The GRU network based solution performed well and won the SentiRuEval-2016 competition on both domains (banks and telecommunication companies).

Using word2vec vectors as features has made a major contribution to the result. However, we believe that parameters of our classifiers were not optimal, even for GRU network. After publication of labeled test data by organizers of the competition, we were able to achieve macro-averaged F1-score above 0.6 on test data for both domains using GRU network. One of the parts of our future work is to find optimal architectures and learning parameters for RNN and CNN. It is also possible to combine RNN and CNN into one compound network.

In addition, our future research includes adapting our neural network based approaches to object-oriented sentiment analysis, as well as developing methods of domain adaptation within these approaches.

Acknowledgements

This work was supported by the Russian Foundation for Basic Research grant 15-37-20375.

References

1. *Adamson A., Turan V. D.,* (2015), Opinion Tagging Using Deep Recurrent Nets with GRUs, available at: https://cs224d.stanford.edu/reports/AdamsonAlex.pdf
2. *Cho K., van Merrienboer B., Gulcehre C., Bougares F., Schwenk H., Bengio Y.,* (2014), Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, CoRR, available at: http://arxiv.org/abs/1406.1078
3. *Chung J., Gulcehre C., Cho K., Bengio Y.,* (2014), Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, CoRR, available at: http://arxiv.org/abs/1412.3555

4. *Collobert R., Weston J., Bottou L., Karlen M., Kavukcuoglu K., Kuksa P.,* (2011), Natural language processing (almost) from scratch, CoRR, available at: http://arxiv.org/abs/1103.0398
5. *dos Santos C., Maira G.,* (2014), Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts, Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, pp. 69–78
6. *Graves A.,* (2012), Supervised Sequence Labelling with Recurrent Neural Networks, available at: http://dx.doi.org/10.1007/978-3-642-24797-2
7. *Hochreiter S., Schmidhuber J.,* (1997), Long Short-Term Memory, Neural computation, volume 9, number 8, pp. 1735–1780
8. *Irsoy O., Cardie C.,* (2014), Opinion Mining with Deep Recurrent Neural Networks, Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, pp. 720–728
9. *Jiang J.,* (2008), Domain Adaptation in Natural Language Processing, available at: http://hdl.handle.net/2142/11465
10. *Jozefowicz R., Zaremba W., Sutskever I.,* (2015), An Empirical Exploration of Recurrent Network Architectures, Proceedings of the 32nd International Conference on Machine Learning (ICML-15), Lille, France, pp. 2342–2350
11. *Kalchbrenner N., Grefenstette E., Blunsom P.,* (2014), A Convolutional Neural Network for Modelling Sentences, CoRR, available at: http://arxiv.org/abs/1404.2188
12. *Kim Y.,* (2014), Convolutional Neural Networks for Sentence Classification, CoRR, available at: http://arxiv.org/abs/1408.5882
13. *Kingma D. P., Ba J.,* (2014), Adam: A Method for Stochastic Optimization, CoRR, available at: http://arxiv.org/abs/1412.6980
14. *Mikolov T., Chen K., Corrado G., Dean J.,* (2013a), Efficient Estimation of Word Representations in Vector Space, CoRR, available at: http://arxiv.org/abs/1301.3781
15. *Mikolov T., Sutskever I., Chen K., Corrado G., Dean J.,* (2013b), Distributed Representations of Words and Phrases and Their Compositionality, CoRR, available at: http://arxiv.org/abs/1310.4546
16. *Mikolov T., Yih W., Zweig G.,* (2013c), Linguistic Regularities in Continuous Space Word Representations, Proceedings of NAACL HLT 2013, Atlanta, USA, pp. 746–751
17. *Palangi H., Deng L., Shen Y., Gao J., He X., Chen J., Song X., Ward R. K.,* (2015), Deep Sentence Embedding Using the Long Short Term Memory Network: Analysis and Application to Information Retrieval, CoRR, available at: http://arxiv.org/abs/1502.06922
18. *Sutskever I., Vinyals O., Le Q. V.,* (2014), Sequence to Sequence Learning with Neural Networks, CoRR, available at: http://arxiv.org/abs/1409.3215
19. *Tang D., Qin B., Liu T.,* (2015), Document Modeling with Gated Recurrent Neural Network for Sentiment Classification, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, pp. 1422–1432

LINGUISTIC DISFLUENCY IN CHILDREN DISCOURSE: LANGUAGE LIMITATIONS OR EXECUTIVE STRATEGY?

Balčiūnienė I. (i.balciuniene@hmf.vdu.lt)[1,2],
Kornev A. N. (k1949@yandex.ru)[1]

[1]Vytautas Magnus University, Kaunas, Lithuania

[2]Saint-Petersburg State Pediatric Medical University, Saint-Petersburg, Russia

The paper deals with linguistic disfluencies (hesitations, repetitions, revisions, false starts, and incomplete utterances) in Russian-speaking language-impaired (N=12) vs. typically-developing (N=12) preschoolers. The corpus-based study aimed at evaluation and comparison of linguistic disfluency in narrative vs. dialogue discourse within and between the groups. Following the *Russian Assessment Instrument for Narrative* (RAIN) methodology, each subject performed two tasks, i.e. storytelling and story retelling according wordless picture sequences; each of the tasks was followed by a structured dialogue based on ten comprehension questions. Both narratives and dialogues were transcribed and annotated for automatized linguistic analysis. Finally, individual measures (a number of each category of disfluencies per utterance) were estimated and submitted for statistical analysis.

Results of our study evidenced that mainly linguistic disfluencies are caused by distinct strategies of speech production due to a level of the subject's language competence, cognitive resource, and the circumstances of narrative and dialogue production.

Keywords: narrative, conversation, corpus-based data, linguistic disfluency

НАРУШЕНИЯ ПЛАВНОСТИ РЕЧЕВОГО АКТА У ДЕТЕЙ: РЕЧЕВАЯ ДИСФУНКЦИЯ ИЛИ СТРАТЕГИЯ АУТОМОНИТОРИНГА?

Балчюниене И. (i.balciuniene@hmf.vdu.lt)[1,2],
Корнев А. Н. (k1949@yandex.ru)[1]

[1]Санкт-Петербургский государственный педиатрический медицинский университет, Санкт-Петербург, Россия

[2]Университет Vytautas Magnus, Каунас, Литва

Ключевые слова: нарратив, разговор, корпусные данные, нарушение плавности речи

1. Introduction

Speech disfluency can generally be distinguished as being either stuttering or linguistic disfluency. The second one contains hesitations, repetitions, revisions, false starts, incomplete utterances, etc. (Loban, 1976; MacLachlan, Chapman, 1988). Linguistic disfluencies seem to be quite natural elements of spontaneous speech in adults (Akxutina, 1989; Fromkin, 1971; Garrett, 1980; Bock, Levelt, 1994; Kibrik, Podlesskaya, 2007) and typically developing (TD) children (Culatta, Leeper, 1989–1990). Some data evidenced that the age might influence a number of disfluencies: its amount increases along the speaker's age (Evans, 1985; Bortfeld et al., 2001; Gyarmathy, Neuberger, 2013) and growing linguistic skills (Fiestas et al., 2005). As Starkweather (1987) has stated, with developing syntax and increasing vocabulary, utterances become longer and linguistically (structurally and semantically) more complex, theoretically making it more difficult to plan a speech. However, here are some opposite observations that child discourse might be more disfluent than the adult one (Garmash, 1999; McDaniel et al., 2010a; 2010b).

Linguistic disfluency has been investigated in various atypically-developing (Culatta, Leeper, 1989–1990; Redmond, 2004; Sieff, Hooyman, 2006; Engelhardt et al., 2010; Steinberg et al., 2013) and specifically language-impaired (SLI)[1] populations (Boscolo et al., 2002; Madon, 2007; Gou et al., 2008). Results of the studies have highlighted that children with language and learning difficulties tend to use more linguistic disfluencies than do their TD peers (Guo et al., 2008). To sum up, a great number of linguistic disfluencies might be a symptom of atypical language acquisition; on the other hand, production of linguistic disfluencies might be treated as natural component of non-prepared, spontaneous discourse (Schegloff et al., 1977; Akxutina, 1989; Fromkin, 1971; Garrett, 1980; Levelt, 1983; 1984; Bock, Levelt, 1994; Postma, 2000; Pillai 2002; Bekenstein, Simpson, 2003; Fox Tree, 2007a; Kibrik, Podlesskaya, 2007; Hennemann, 2015), where a self-monitoring performs a crucial role.

Despite numerous studies, substantially a nature and mechanisms of linguistic disfluency however remain unclear both in children and adults. According psycholinguistic approach, linguistic disfluency might be recognized as a concordance problem between utterance programming, related speech production and on-line self-monitoring (Gyarmathy, Neuberger, 2013), different underlying production problems or strategies for correcting problems. In some studies, a correlation between linguistic disfluency production, individual intelligence, and executive function (Engelhardt et al., 2010; 2013) has been revealed. A number of studies in child language (MacLachlan, Chapman, 1988; Leadholm, Miller, 1992; Wagner et al., 2000; Madon 2007) have confirmed that linguistics disfluencies are more numerous in a narrative than in a conversation, because the narrative context is usually more demanding than the conversational one.

It should be noted that the majority of studies have been focused on just particular type of linguistic disfluencies, e.g. hesitations (Fox Tree, 2007a; 2007b; Corley,

---

[1] In Russian logopaedics, specific language impairment (SLI) in children usually is named as the *primary speech and language disorder*.

Stewart, 2008) or revisions and self-repairs (Levelt, 1983; 1984; Evans 1985; Bekenstein, Simpson, 2003). In this paper, we aim at discussing linguistic disfluencies as the entity of various kinds of disruptions of linguistically fluent speech. In some studies, an influence of the discourse genre on a frequency of dysfluencies has been found. A narrative might be recognized as a high cognitive loading type of discourse, while a dialog seems less planning demanded. In our corpus-based study, these discourse types were compared from the perspective of disfluency production.

Research questions addressed in the paper:

1. How does a frequency and distribution of linguistic disfluencies distinct between typically developing (TD) and specifically language-impaired (SLI) children?
2. How does a frequency and distribution of linguistic disfluencies distinct between different discourse genres, i.e. child narrative vs. structured dialogue with an adult?

2. Types of Linguistic Disfluencies

In our data, disfluencies were grouped into hesitations, repeats, revisions, false starts, and incomplete utterances; all disfluencies were coded with special symbols according internationally accepted principles of discourse annotation (MacWhinney, 2010), as exemplified below.

2.1. Hesitations

Hesitations can be described as *silent* (*unfilled*) or *filled pauses* (*fillers*) involving an articulation of some sound(s) during the delay (Watanabe, Rose, 2012). Silent pauses (1, 2) can be defined as periods of silence longer than the pauses in an equivalent fluent utterance (Fraundorf, Watson, 2013), e.g.:

(1) *I lisa* (.) *uvidela i zaxvatila za lapu*.
'And the fox [PAUSE] saw [it] and grabbed [his] foot.'

(2) *Malen'kij kozleno*(.)*k zaxotel po*(.)*kupats'a*.
'The baby-goat-[PAUSE within the word] wanted to swim-[PAUSE within the word].'

Fillers can be defined as verbal interruptions that do not relate to the proposition of the main message (Fraundorf, Watson, 2013), and they can be further classified into (3, 4) nonlexical formations (such as *uh* and *um* in English) and (5, 6) semantically insignificant words and constructions (such as *well, like*, *you know* in English), e.g.:

(3) *Potom sobaka* (.) *mmm* (.) *prognala koshku*.
'Then the dog [PAUSE-FILLER-PAUSE] chased the cat away.'

(4) *I uvidela, chto kot* (.) *aaa* (.) *xochet skushat' ptenchikov.*
'And [the dog] saw that the cat [PAUSE-FILLER-PAUSE] wanted to eat the baby-birds.'

(5) *Sobaka za nej poshla. To est', pobezhala.*
'The dog went to her. I mean, [it] run.'

(6) *Ptica uletela za cherv'achkami. Za cherv'achkami. Vobshchem, kormit'.*
'The mother-bird flied [to look] for worms. For worms. Like, to feed [the baby-birds].'

2.2. Repeats

Repeats (i.e., unmodified repetitions) can be grouped into repeated (7, 8) parts of a word, (9, 10) words, and (11, 12) strings of words, e.g.:

(7) *Maly—* (.) *malyshi ptency ostalis' odni.*
'Baby:INCOMPLETE-PAUSE-baby-birds were left alone.'

(8) *A lisa xotela by s'e—* (.) *s'est'.*
'And the fox would want to eat: INCOMPLETE-PAUSE-to eat.'

(9) *I* (.) *i ona uvidela, chto ptency spaseny.*
'And [PAUSE] and she saw that the baby-birds were saved.'

(10) *Ona* (.) *ona zalezla na derevo i nachala podbirats'a k* (.) *k ptenchikam.*
'She [PAUSE] she climbed up the tree and started sneaking to [PAUSE] to the chicks.'

(11) *I skazali* (.) *i skazali èto* (.) *svojej mame.*
'Then [they] told [PAUSE] then [they] told this [PAUSE] to their mother.'

(12) *A lisa smotrela za kozlenkom za kozlenkom.*
'And the fox was looking at the baby-goat at the baby-goat.'

2.3. Revisions

Revision (or *repairs*) are self-corrections of material already spoken (Fraundorf, Watson, 2013) and they can be classified into (13, 14) phonological, (15, 16) lexical, and (17, 18) grammatical modifications of speech.

(13) *Prinesla odnogo cherv'aka dra* (.) *dl'a vsex ptenchikov.*
'[She] brought just one worm for:INCORRECT for all the chicks.'

(14) *Esli by ne xvatila (.) ne sxatila ee za xvost.*
'If [she] [did] not grab [PAUSE] [did] not grab her tail.'

(15) *A ptica, kotoraja na suku lezhala nu (.) sidela, èto vse zametila.*
'And the bird who [was] laying [FILLER-PAUSE] [was] sitting on the branch saw all this.'

(16) *Xotela s'est' pti— nu (.) lisu.*
'[The bird] wanted to eat the bird:INCOMPLETE-FILLER the fox.'

(17) *Byli (.) bylo mama ptica i tri ptenca.*
'[There] were-PAUSE-was a mother-bird and three chicks.'

(18) *Ona ej skazala: nikogda (.) uxodi i nikogda ne prixodi*[2].
'She said to her: never-PAUSE-go away and never come back.'

A process of revision presumes that a speaker repairs previously uttered something erroneous (Fraundorf, Watson, 2013); however, sometimes an incorrect utterance might be replaced by another incorrect utterance (19), also, a correct utterance might be reformulated into (20) another correct utterance and even into (21) an erroneous utterance (Balčiūnienė, 2013), e.g.:

(19) *A potom prishel lis i sxvatil baran-- nu jagnenka*[3].
'And then the fox came and grabbed the sheep:INCOMPLETE-FILLER the lamb.'

(20) *Ona poletela z— eshche za edoj.*
'She flied [to look] for-INCOMPLETE again for food.'

(21) *A lisichka ubexhala (.) uletela.*
'And the fox run [PAUSE] flied away.'

2.4. False starts

False starts (22, 23) can be equaled to crossings out in writing, i.e. a speaker starts producing an utterance and drops it after just a few words, e.g.:

(22) *Kogda na beregu… Oj, ne znaju dazhe, kak ètot malysh nazyvaets'a.*
'When ashore… Oh, I do not know even how to call this baby.'

---

[2] Syntactic reformulation. More on this see Kilani-Schoch et al. (2008), Kazakovskaya, Balčiūnienė (2012a, 2012b).

[3] There is a baby-goat but not a lamb on the picture.

(23) *I potom mama prishla i pomogla emu vybratsa. A on mmm... I lisa uvidela i zaxvatila za lapku.*
'And then the mother came and helped him to get out. But he [FILLER]… And the fox saw [him] and grabbed [his] foot.'

We recognize false starts as separate type of linguistic disfluencies, which, in contrast to the revisions and repetitions, are neither revised nor repeated after dropping them.

2.5. Incomplete (abandoned) utterances

Incomplete (abandoned) utterances (24, 25) considered any utterances where the obligatory ending is missing, e.g.:

(24) *Ona zalezla na derevo i… Zalezla na derevo i obliznulas'.*
'She climbed up the tree and… [She] climbed up the tree and licked [her lips].'

(25) *Oni by skazali, chto koshka za nimi... A sobaka ix spasla.*
'They would tell, that the cat… But the dog saved them.'

3. Research Method

3.1. Participants

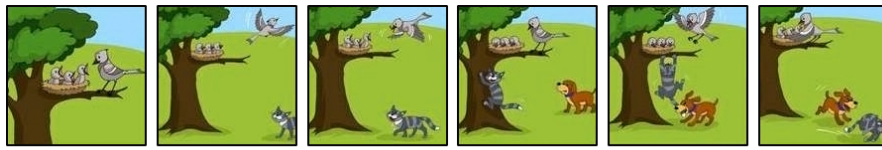
The data contains a part of the *Corpus of Russian TD and SLI children language* (Kornev et al., 2015) available online at http://rclc.iling.spb.ru/corp/[4]. The subjects of the study were 12 clinically-referred monolingual 6-year old (mean age 76 months) SLI children who received 2 years course of speech therapy and 12 TD peers. SLI children were recruited from those who attended remedial treatment unit for speech and language disordered kindergartens. Exclusion criterion was non-verbal IQ on Raven's matrix below 24. In all cases, morphosyntactic backwardness (below 5 year level) was coupled with articulation/phonological disorders. TD children were recruited from day care center for kindergartens. For both the TD and SLI group, informed consent was obtained from parent before the experiment.

[4] The Corpus was developed in the framework of a national project *Development of linguistic sub-systems in typically-developing and language-impaired children: Corpus-based and experimental study* [Formirovanie jazykovyh podsistem u detej s normoj i otstavaniem v razvitii rechi: korpusnoe i èksperimental'noe issledovanie tekstov] carried out with the financial support of the Russian Foundation for Humanities, grant No. 14-04-00509.

3.2. Task Administration

The subjects were assessed by means of the *Russian Assessment Instrument for Narratives—RAIN* (Kornev, Balčiūnienė, 2014; 2015). After warming up, each subject performed two tasks, i.e. storytelling and story retelling according wordless picture sequences *the Baby-Birds* and *the Baby-Goats* (see Figure 1); each of the tasks was followed by ten comprehension questions (CQ) (see Table 1).

The *Baby-Birds* sequence



The *Baby-Goats* sequence

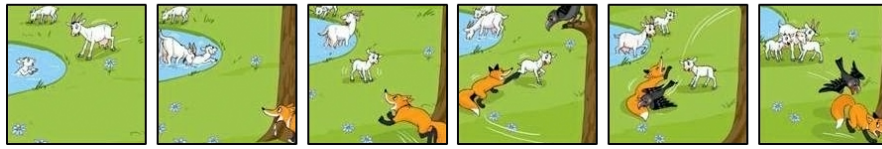


Fig. 1. Stimulus material (based on Gagarina et al., 2012)

Table 1. Comprehension questions (based on Gagarina et al., 2012)

The Baby-Birds	***The Baby-Goats***
0. Did you like the story?	0. Did you like the story?
1. *Why do the baby-birds open their mouths?*	1. *Why does the baby-goat shout?*
2. *Why did the mother-bird fly away?*	2. *Why did the mother-goat got to the water?*
3. *Why did the cat climb the tree?*	3. *Why did the fox grab leap to the baby-goat?*
4. *Why did the dog grab the cat's tail?*	4. *Why did the bird grab the fox's tail?*
5. *Why did the dog chase the cat?*	5. *Why did the bird chase the cat?*
6. *What was the dog thinking when he was chasing the cat?*	6. *What was the bird thinking when she was chasing the fox?*
7. *What was the cat thinking when she was being chased by the cat?*	7. *What was the fox thinking when she was being chased by the bird?*
8. *If the mother-bird was able to speak, what would she say to the dog?*	8. *If the mother-goat was able to speak, what would she say to the bird?*
9. *If the mother-bird was able to speak, what would she say to the cat?*	9. *If the mother-goat was able to speak, what would she say to the fox?*
10. *If the baby-birds were able to speak, what would they say to their mother when she came back?*	10. *If the second baby-goat was able to speak, what would he say to his mother when she came to him?*

The sessions of the 1st and the 2nd task were separated by a few minutes of free talk between the interviewer and the child; the order of tasks was counterbalanced (see Table 2) with regard to narrative mode (telling vs. retelling) and story complexity (the *Baby-Goats* considered more complex than the *Baby-Birds*; see Kornev, Balčiūnienė, 2014; 2015).

Table 2. Counterbalancing scheme

Children No.	Session No. 1	Session No. 2
1, 5, 9	Telling *Baby-Birds*	Retelling *Baby-Goats*
2, 6, 10	Telling *Baby-Goats*	Retelling *Baby-Birds*
3, 7, 11	Retelling *Baby-Birds*	Telling *Baby-Goats*
4, 8, 12	Retelling *Baby-Goats*	Telling *Baby-Birds*

3.3. Data Coding and Analysis

All the stories and dialogues were video-recorded and transcribed according to the conventions of the *Codes for the Human Analysis of Transcripts* (CHAT; MacWhinney, 2010) by three graduate transcribers. They were blind to the diagnosis of the participants. During discourse annotation, the main structural characteristics, such as turn-takings, overlaps, pauses, repeats, revisions, etc., were encoded for the analysis. Then, individual measures (a number of each category of disfluencies per utterance) were estimated and submitted for statistical analysis by means of one way ANOVA and paired-samples T-test.

4. Results

Due to our aim at comparing disfluencies in different discourse genres (narrative and structured dialog), the data analysis was performed separately for storytelling and for dialog based on the CQ. The total number of linguistic disfluencies per utterance produced in the story-telling was the same both in the SLI and the TD groups (see Table 3).

However, some forms of disfluencies discriminated the groups (see Table 3). Among all hesitations, the filled hesitations (fillers) were more dominant (72%) in the TD children, whereas the unfilled hesitations were numerous (54%) in the SLI peers. The incomplete utterances were significantly more numerous in the SLI children.

The total number of repetitions was similar within both the groups, but, again, some forms of repetitions discriminated the groups. The repeated parts of word were significantly prevalent among all repeats in the SLI children while repeated words were the most frequent in the TD peers. The total amount of revisions did not differ between the groups but the phonological revisions were significantly more numerous in the TD children than in the SLI peers (see Table 3).

Table 3. Distribution of linguistic disfluencies in the storytelling within the SLI vs. TD groups

	SLI N = 12		**TD N = 12**				
Measures	***M***	***σ***	***M***	***σ***	***df***	***F***	***Sig.***
1. A total number of disfluencies per utterance	0.88	0.43	0.63	0.42	1	0.8	0.40
2. A number of false starts per utterance	0.008	0.02	0.064	0.086	1	3.19	0.092
3. A number of incomplete utterances per utterance	0.09	0.07	0.027	0.054	1	4.45	0.05
4. A number of hesitations per utterance	0.40	0.26	0.47	0.26	1	0.12	0.74
4.1. A percentage of filled hesitations among all hesitations	0.34	0.33	0.72	0.20	1	9.34	0.007
4.2. A percentage of unfilled hesitations among all hesitations	0.64	0.379	0.28	0.20	1	3.63	0.07
5. A number of repetitions per utterance	0.12	0.06	0.36	0.33	1	1.48	0.27
5.1. A percentage of repeated parts of word among all repeats	0.40	0.46	0.01	0.04	1	8.24	0.01
5.2. A percentage of repeated words among all repeats	0.31	0.44	0.716	0.46	1	3.74	0.07
6. A number of revisions per utterance	0.07	0.06	0.18	0.24	1	0.56	0.48
6.1. A percentage of phonetical revisions among all revisions	0.10	0.20	0.48	0.43	1	5.09	0.04

Comparative paired-samples T-tests statistical analysis of disfluencies in storytelling vs. dialogue discourse revealed significant differences in the SLI group (means 0.77 and 0.50 respectively; $t = 2.4$; $P \leq 0.03$) and in the TD group (0.82 and 0.45 respectively; $t = 2.36$; $P \leq 0.05$).

The children produced more filled hesitations in the narratives than in the dialogues (means 0.22 and 0.12 respectively; $t = 1.93$; $P \leq 0.069$); they also produced more revisions (per utterance) in total (means 0.16 and 0.07 respectively; $t = 2.48$; $P \leq 0.023$); finally, the phonological revisions were produced in only narrative discourse (means 0.00 and 0.07 respectively; $t = 3.29$; $P \leq 0.004$).

Dynamic approach to narrative analysis elaborated in our previous studies (Kornev, Balčiūnienė, 2014; 2015) for clinical practice evidenced that narrative production process depended on variables such as task order, story complexity and mode. However, the general linear model of statistical analysis did not reveal any significant dynamic influence of the story cognitive complexity and task order on the production of disfluencies.

5. Discussion and Conclusions

Child language development might be treated from twofold perspective: from the linguistic structural static dimension and from psycholinguistic on-line discourse analysis dimension. The last one focuses on the process of planning, programming, and self-monitoring. Linguistic disfluencies have close relations to all of these processes and thus they should be appropriate basis for studying the processes of discourse production in children.

In our study, we addressed some questions permanently debated in many previous publications related to adult spontaneous (unprepared) discourse (e.g., Levelt, 1983; 1984; Kibrik, Podlesskaya, 2007). It is interesting to cite some assumptions that fillers might play not the same functional role in the child speech as in the adult one (Pepinsky et al., 2001; Taelman et al., 2009).

The main question addressed the nature of linguistic disfluency: is it some kind of failure to concord speech programming and production processes, a manifestation of child language immaturity or a complex of individual strategies related to speech self-monitoring used by the narrator in a discourse production. The total number of disfluencies was rather close in both groups and thus this did not support the "immaturity" hypothesis. But some qualitative differences between the groups were observed. For example, when faced with some problems of the utterance programming, the TD children tended to use filled hesitations, whereas the SLI children used to explore silent pauses or to repeat part of a word. Besides that, language impaired subjects tended to produce shorter utterances ($MLU_{SLI} = 4.17$; $MLU_{TD} = 5.62$; $F = 6.37$, $P \leq 0.040$) to reduce the cognitive loading in utterance programming.

The findings encourage discussing child speech disfluencies as the on-line strategies to resolve complications in utterance programming in parallel with speech production. It looks obvious that usually the SLI children start uttering before they finished elaborating the plan of the utterance and consequently perform its language programming in parallel with speech execution. Their limited cognitive resources overloading and thus prevent completing the utterance. In the case of child-adult structural dialog cognitive loading is reduced and disfluencies appear less frequently. To sum up, in childhood linguistic disfluencies represent the complex of distinct strategies in discourse production due to a level of the subject's language competence, cognitive resource, and the circumstances of narrative production.

References

1. *Akxutina T. V.* (1989), Language Production. Neurolinguistic Syntactic Analysis [Porozhdenije rechi. Neirolingvisticheskij analiz sintaksisa], Moscow.
2. *Balčiūnienė I.* (2013), Linguistic disfluency in narrative speech: Evidence from story-telling in 6-year olds, INTERSPEECH-2013, pp. 2143–2146.
3. *Bekenstein R., Simpson A. P.* (2003), Phonetic correlates of self-repair involving word repetition in German spontaneous speech, in R. Englund (Ed.) Gothenburg Papers in Theoretical Linguistics, Vol. 90, pp. 81–84.

4. *Bock K., Levelt W. J.M.* (1994), Language production. Grammatical encoding, in M. A. Gernsbacher (Ed.) Handbook of Psycholinguistics, Academic Press, New York, pp. 741–779.
5. *Bortfeld H., Leon S. D., Bloom J. E., Schober M. F., Brennan S. E.* (2001), Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender, Language and Speech, Vol. 44 (2), pp. 123–147.
6. *Boscolo B., Bernstein Ratner N., Rescorla L.* (2002), Fluency of school-aged children with a history of specific expressive language impairment: An exploratory study, American Journal of Speech-Language Pathology, Vol. 11, pp. 41–49.
7. *Corley M., Stewart O. W.* (2008), Hesitation disfluencies in spontaneous speech: The meaning of *um*, Language and Linguistics Compass, Vol. 4, pp. 589–602.
8. *Culatta R., Leeper L. H.* (1989–1990), The differential diagnosis of disfluency, National Student Speech Language Hearing Association Journal, Vol. 17, pp. 59–64.
9. *Engelhardt P. E., Corley M., Nigg J. T., Ferreira F.* (2010), The role of inhibition in the production of disfluencies, Memory & Cognition, Vol. 38 (5), pp. 617–628.
10. *Engelhardt P. E., Nigg J. T., Ferreira F.* (2013), Is the fluency of language outputs related to individual differences in intelligence and executive function? Acta Psychologica (Amst), Vol. 144 (2), 424, pp. 432.
11. *Evans M. A.* (1985), Self-initiated speech repairs: A reflection of communicative monitoring in young children, Developmental Psychology, Vol. 21 (2), pp. 365–371.
12. *Fiestas Ch E., Bedore L. M., Peña E. D., Nagy V. J.* (2005), Use of mazes in the narrative language samples of bilingual and monolingual 4- to 7-year old children, in J. Cohen, K. T. McAlister, K. Rolstad, J. MacSwan (Eds.) Proceedings of the 4th International Symposium on Bilingualism, Cascadilla Press, Somerville, MA, pp. 730–740.
13. *Fox Tree J. E.* (2007a), Functional spontaneous speech phenomena, Perspectives on Fluency and Fluency Disorders, Vol 17 (2), pp. 17–19.
14. *Fox Tree J. E.* (2007b), Folk notions of um and uh, you know, and like, Text & Talk, Vol. 27 (3), pp. 297–314.
15. *Fraundorf S. H., Watson D. G.* (2013), Alice's adventures in *um*-derland: Psycholinguistic sources of variation in disfluency production, Language and Cognitive Processes, Vol. 29 (9), pp. 1083–1096.
16. *Fromkin V.* (1971), The non-anomalous nature of anomalous utterances, Language, Vol. 47 (1), pp. 27–52.
17. *Gagarina N., Klop D., Kunnari S., Tantele K., Välimaa T., Balčiūnienė I., Bohnacker U., Walters J.* (2012), MAIN: Multilingual Assessment Instrument for Narratives, ZAS, Berlin.
18. *Garmash N. G.* (1999), An Influence of Hesitations on a Production of Child Oral Discourse: Oral Retellings and Spontaneous Speech in 4–9 yers age children [Vlijanije xezitaciji na organizaciju ustnogo detskogo diskursa (na materiale ustnyx pereskazov i spontannoj rechi detej 4–9 let)], Moscow State University, Moscow.
19. *Garrett M. F.* (1980), The limits of accommodation, in V. Fromkin (Ed.) Errors in Linguistic Performance, Academic, New York, pp. 263–271.

20. *Guo L., Tomblin J. B., Samelson V.* (2008), Speech disruptions in the narratives of English-speaking children with specific language impairment, Journal of Speech, Language, and Hearing Research, Vol. 51 (3), pp. 722–738.
21. *Gyarmathy D., Neuberger T.* (2013), Self-monitoring strategies: the factor of age, Paper presented at the 19th International Congress of Linguists, July 21–27, Geneve, Switzerland.
22. *Hayward D., Schneider Ph.* (2006), Effectiveness of teaching story grammar knowledge to pre-school children with language impairment: An exploratory study, Child Language Teaching and Therapy, Vol. 16 (3), pp. 255–284.
23. *Hennemann A.* (2015), The phenomenon of self-repair in Spanish and Portuguese, Procedia—Social and Behavioral Sciences, Vol. 173, pp. 279–284.
24. *Kazakovskaya V. V., Balčiūnienė I.* (2012a), Interrogatives in Russian and Lithuanian motherese: Do we communicate with our children in the same way? The Journal of Baltic Studies, Vol. 43 (2), pp. 197–218.
25. *Kazakovskaya V. V., Balčiūnienė I.* (2012b), Lithuanian and Russian child-directed speech: Why do we ask young children so many questions? Estonian Papers in Applied Linguistics, Vol 8, pp. 69–89.
26. *Kibrik A. A., Podlesskaya V. I.* (2007), Self-repairs and other types of linguistic disfluencies as a target for annotation of spoken language corpora [Samoispravlenija govorjashchego i drugije tipy rechevyx sbojev kak objekt annotirovanija v korpusax ustnoj rechi], Scientific and Technological Information, Vol. 2: Processes and Systems of Information [Nauchno-texnicheskaya informacija. Serija 2: Informacionnyje processy i sistemy], pp. 2–23.
27. *Kilani-Schoch M., Balčiūnienė I., Korecky-Kröll K., Laaha S., Dressler W. U.* (2008), On the role of pragmatics in child-directed speech for the acquisition of verb morphology, Journal of Pragmatics, Vol. 41 (2), pp. 129–159.
28. *Kornev A. N., Balčiūnienė I.* (2014), Story (re-)telling and reading in children with dyslexia: Language or cognitive resource deficit? In Book of Abstracts: LSCD-2014. UCL, London, pp. 23–26.
29. *Kornev A. N., Balčiūnienė I.* (2015), Narrative production weakness in Russian dyslexics: Linguistic or procedural limitations? Estonian Papers in Applied Linguistics, Vol. 11, pp. 141–157.
30. *Kornev A. N., Balčiūnienė I., Voeikova M. D., Ivanova K. A., Yagunova E. V.* (2015), Language acquisition in typically developing and language impaired children: A corpus-based analysis of spontaneous and elicited texts [Formirovanije jazyka u detej s normoj i otstavanijem v razvitiji rechi. Korpusnoje issledovanije spontannyx i vyzvannyx tekstov]. In N. N. Kazanskyj, M. D. Voeikova, E. G. Sosnovceva (Eds.) Acta Linguistica Petropolitana. ILI RAN, Saint-Petersburg, pp. 605–624.
31. *Leadholm B. J., Miller J. F.* (1992), Language Sample Analysis: The Wisconsin Guide, Wisconsin Department of Public Instruction, Madison, WI.
32. *Levelt W.* (1983), Monitoring and self-repair in speech, Cognition, Vol. 14, pp. 41–104.
33. *Levelt W.* (1984), Spontaneous self-repairs in speech: Processes and representations, in M. P. R. Van den Broecke, A. Cohen (Eds.) Proceedings of the Tenth International Congress of Phonetic Sciences, Foris Publications Dordrecht, Holland.

34. *Loban W.* (1976), Language Development: Kindergarten through grade twelve. National Council of Teachers of English, Urbana, Ill.
35. *MacLachlan B.G., Chapman R. S.* (1988), Communication breakdowns in normal and language learning-disabled children's conversation and narration, Journal of Speech and Hearing Disorders, Vol. 53, pp. 2–7.
36. *MacWhinney B.* (2010), The CHILDES Project: Tools for Analyzing Talk. Electronic Edition. http://childes.psy.cmu.edu/manuals/CLAN.pdf.
37. *Madon Z.* (2007), Investigation of Maze Production in Children with Specific Language Impairment. MA thesis. McGill University, Montreal.
38. *McDaniel D., McKee C., Garret M. F.* (2010a), Children›s sentence planning: Syntactic correlates of fluency variations, Journal of Child Language, Vol. 37 (1), pp. 59–94.
39. *McDaniel D., McKee C., Garret M. F.* (2010b), Fluency markers for children›s sentence planning: Early and late stage processing. In Proceedings of the 35th annual Boston University Conference on Language Development. Cascadilla Press, Somerville, MA.
40. *Pepinsky Th., Demuth K., Roark B.* (2001), The status of filler syllables in children's early speech. In Proceedings of the 25th annual Boston University Conference on Language Development. Cascadilla Press, Somerville, MA.
41. *Pillai S.* (2002), Error-detection, self-monitoring and self-repair in speech production, Proceedings of the 9th Australian International Conference on Speech Science & Technology, Melbourne, December 2 to 5, 2002, pp. 533–537.
42. *Postma A.* (2000), Detection of errors during speech production: A review of speech monitoring models, Cognition, Vol. 77, pp. 97–131.
43. *Redmond S. M.* (2004), Conversational profiles of children with ADHD, SLI and typical development, Clinical linguistics & Phonetics, Vol 18 (2), pp. 107–125.
44. *Schegloff E. A., Jefferson, G., Sacks H.* (1977), Preference for self-correction in the organization of repair in conversation, Language, Vol. 53 (2), pp. 361–382.
45. *Sieff, Sh., Hooyman B.* (2006). Language-based disfluency: A child case study. In: ASHA 2006 Convention, available at: https://www.mnsu.edu
46. *Starkweather C. W.* (1987), Fluency and Stuttering, Prentice Hall, Englewood Cliffs, NJ.
47. *Steinberg M. A., Bernstein Ratner N., Gaillard, W., Berl M.* (2013), Fluency patterns in narratives from children with localization related epilepsy, Journal of Fluency Disorders, Vol. 38 (2), pp. 193–205.
48. *Taelman H., Durieux G., Gillis S.* (2009), Fillers as signs of distributional learning, Journal of Child Language, Vol. 36, pp. 323–353.
49. *Wagner C. R., Nettelbladt U., Sahlén U., Nilholm C.* (2000), Conversation versus narration in preschool children with language impairment, International Journal of Language and Communication Disorders, Vol. 35, pp. 83–93.

О ДИСКУРСИВНЫХ РЕЖИМАХ ИСПОЛЬЗОВАНИЯ ОЦЕНОЧНЫХ СЛОВ И ВЫРАЖЕНИЙ

Баранов А. Н. (baranov_anatoly@hotmail.com)

Институт русского языка РАН, НИУ Высшая школа экономики, Москва, Россия

Ключевые слова: языковая оценка, дескрипция, дискурсивный режим, лингвистическая экспертиза

ABOUT DISCURSIVE MODES OF EVALUATION IN RUSSIAN

Baranov A. N. (baranov_anatoly@hotmail.com)

Institute of Russian Language (Vinogradov's Institute), National Research University Higher School of Economics, Moscow, Russia

The paper discusses different modes of evaluation in Russian. Evaluation is considered as a speech act based on a cognitive procedure which has the following form: (i) evaluation of an object *X* as possessing a feature *q* consists of comparing of parameter *Q* with *X* and picking out of *q* as a function of *Q* with an argument of *X*; (ii) the feature *q* presupposes recommendations for decision making in connection with an object *X*. Cognitive procedure of description of an object X as possessing of a feature *q* doesn't presupposes any recommendation for decision making.
In some discursive modes semantics of evaluation lose its influence force or at least it getting weaker. Discursive mode is defined as a sphere of functioning of speech forms in discourse, in which their meaning regularly changed. Different discourses allow different kinds of discursive modes. In the paper are discussed the following discursive modes, which modify evaluation force: irony, language game, common nomination, indefinite reference.

Key words: language evaluation, description, discursive mode, forensic linguistics

1. Оценка и оценочное суждение

В лингвистике оценка исследовалась в первую очередь в лексической семантике и стилистике. В первом случае обращалось внимание на существование слов, в семантике которых присутствует оценочный компонент (*хороший — плохой*; *пресловутый, добряк, негодяй*), а также на оценочные коннотации лексем (ср. лексему *свинья* с коннотацией 'грубый' или лексему *ветер* с коннотацией 'переменчивости' — Апресян 1995). Во втором — оценочность рассматривалась как свойство выразительности речи в целом, как особенность стиля, присущая данному типу дискурса или являющаяся индивидуальной характеристикой говорящего (Телия 1991; Шмелев 1964). Отмечалось, что оценочная семантика свойственна языковым единицам различных уровней (Вольф 1985: 5–7). Лингвистические классификации языковой оценки так или иначе, с теми или иными девиациями повторяли известные философские и логические классификации, разделявшие оценку на общую и частную (Арутюнова 1984). Соответственно, слова *хороший, плохой* выражают общую оценку, а прилагательные *красивый, безобразный, комфортабельный, неудобный, умный, глупый, гениальный, тупой, богатый, бедный* — частную. Лингвистическая теория не дает, однако, ясного ответа на вопрос об оценках, передаваемых в тексте косвенно — без использования оценочной лексики. Кроме того, некоторые феномены оценки, существенные для приложений лингвистики, не имеют очевидных коррелятов в лингвистической семантике.

В лингвистической экспертизе широко используется категория «оценочного суждения», представленная в законодательстве как неопределяемый термин. Эту категорию можно определить как вид утверждения, в пропозиции которого устанавливается связь между объектом оценивания и значением параметра общей оценки ('хорошо — плохо') или значением параметра той или иной частной оценки. Оценочные суждения субъективны и в силу этого их нельзя проверить на соответствие действительности. Например, утверждение *Роман — взяточник* может быть истинным или ложным, а оценочное суждение *Ольга красива* — нет, поскольку красота человека является оценочной субъективной категорией.

Некоторые частные оценки в разных типах дискурсов, обусловленных как исторически, так и социально, могут менять свое соотношение с общими оценками. Так, оценочное прилагательное *богатый* в революционном дискурсе советской эпохи соотносилось с оценкой 'плохо', а *бедный* — с оценкой 'хорошо'. В дореволюционный период, а также в дискурсе современной российской политики *богатый*, наоборот, 'хорошо', а *бедный* — 'плохо'. Такова специфика использования оценочных слов в идеологизированных дискурсах. В этом отношении оценочные суждения близки и к категории мнения.

Следует отметить, что среди частных есть оценки, близкие к дескрипциям (утверждениям о свойствах и признаках). Так, утверждения с параметрическими прилагательными *высокий* и *низкий*, *широкий* и *узкий*, *длинный* и *короткий* во многих ситуациях допускают истинностную интерпретацию. Это связано с тем, что параметрические прилагательные содержат в своей семантике

значительный дескриптивный компонент. Соответственно, фразы типа *Шкаф слишком высок, Потолок в комнате низкий, Коридоры в здании узкие и длинные* существенно ближе по своим свойствам к утверждениям. Кроме того, прилагательные указанного типа нельзя легко соотнести с общей оценкой. Высокий шкаф хорош для одних ситуаций и плох для других. Тем самым, утверждения, апеллирующие к оценкам с существенной дескриптивной частью семантики, не могут рассматриваться как оценочные суждения в точном смысле слова.

Оценка, однако, может выражаться не только в оценочных суждениях, но и в оценочной составляющей слов, входящих в обычные неоценочные утверждения и другие типы речевых актов. Так, Е. М. Вольф отмечает, что не всякое высказывание с семантикой оценки может рассматриваться как речевой акт оценки (Вольф 1985). Кроме того, оценка (как положительная, так и отрицательная) может находиться в разных частях семантики слова и высказывания.

2. Семантика оценки и ее когнитивные корреляты

В лингвистической теории не сформировалась сколь-нибудь цельная теория оценки. Может быть, из-за этого когнитивный аспект оценивания фактически был исключен из рассмотрения. Между тем, оценка прямо связана с процессом принятия решений. В ряде работ по семантике и прагматике оценки было показано, что процедура оценивания является частью процесса принятия решений (Хэар 1985; Баранов 1989). Оценивание (точнее результат оценивания) позволяет выбрать ту альтернативу решения проблемной ситуации, которая в большей степени удовлетворяет желаниям и намерениям человека, делающего тот или иной выбор. С когнитивной и семантической точки зрения оценка несет в себе следы прескрипции: оценка чего-то как хорошего как бы рекомендует использовать это хорошее, а оценка чего-то как плохого, соответственно, влечет рекомендацию отказаться от плохого.

Языковые следы ментальной процедуры оценивания обнаруживаются в первую очередь в области иллокутивной семантики и в сфере лексического значения. В первом случае следует говорить о речевом акте оценки, а во втором — о соотношении оценочных и дескриптивных смыслов в плане содержания слова (проблематика, типичная для лингвистического подхода). Компоненты содержания оценки как особого типа речевого акта определяются структурой соответствующей ментальной процедуры. Говорящий, оценивая некоторый объект *X* как *q* (как обладающий характеристикой *q*), делает следующее: определив некоторый признак *Q* [основание оценки] и, сопоставив *Q* с *X* [объект оценки], выбирает значение признака *Q*, равное *q*, и приписывает *q X*, полагая, что выбор *q* для *X* может влечь некоторое практическое следствие [аналог акта принятия решения]. Признак *Q* в случае общей оценки соответствует оценочной шкале «Хорошо — Плохо».

В существующих классификациях речевых актов оценка не представлена особой категорией. Наиболее близкими к акту оценки можно считать экспрессивы по классификации Дж. Серля, сущность которых состоит в том, чтобы

«выразить психологическое состояние, задаваемое условием искренности относительного положения вещей, определенного в рамках пропозиционального содержания» (Серль 1986: 183). Можно предположить, что к экспрессивам относятся выражения одобрения и неодобрения, комплименты, похвала и многие другие, выражающие субъективную оценку.

В теоретической модели оценивания, предложенной в (Баранов 1989), смысл ‘характеристика’ (*q*) выступает как инвариант дескрипции и оценки. Варьирование оценочных и дескриптивных свойств характеристики *q* определяется эксплицитностью основания оценки (*Q*), с которым связана характеристика *q*, и практического следствия — аналога принятия решения (*Y*). Предположительная связь *q* с *Y* отражает представления говорящего об участии оценки в акте принятия решения, то есть о «рекомендательной» силе оценивания. Чем менее определен признак *Q*, его шкала, выбор *q* и основания этого выбора (то есть чем менее рестриктивна характеристика), чем меньше ясности с принятием решения *Y*, тем меньше собственно оценочных свойств у характеристики *q*. Тем самым, истинная оценка отличается от истинной дескрипции прежде всего (i) **каузальностью** (то есть наличием обоснования характеристики *q* — установлением признака *Q* и сопоставлением его с *X*), (ii) **рестриктивным выбором** (выбор значения признака Q из упорядоченного множества других значений на основании его сопоставления с *X*) и (iii) **ориентацией на акт принятия решения**.

Указанные характеристики акта оценки позволяют объяснить некоторые специфические особенности сочетаемости слов оценочной семантики, предсказывая их собственно оценочные и дескриптивные употребления. Так, вопрос с вопросительным словом *почему* более уместен по отношению к оценке, чем к дескрипции. Действительно, оценка должна быть внутренне обоснована, аргументирована, ср. — *Вы уже работали раньше? — Да, недолго. — Это* ***плохо****. — Почему? — Потому что недолго*, дескрипция же в точном смысле спонтанна и часто неконтролируема, ср. — *Комната была совсем пуста, лишь в середине стоял* ***круглый*** *дубовый стол. —* ?*Почему* ***круглый****?* Отсюда легкость мотивировки оценки и затруднительность мотивировки истинной дескрипции, ср. *Он неосторожен/вероломен, потому что …* и ?*Снег бел, потому что …* Мотивировка дескрипций возможна, однако для этого приходится привлекать законы онтологии, закономерности устройства реального мира, ср. *Снег бел, потому что так устроен мир, Этот стол круглый, потому что его так сделали* и т. п.

Оценочные слова и выражения не функционируют в речи автономно, а входят в контексты, подавляющие или, наоборот, усиливающие оценочную семантику этих речевых форм. Такие контексты можно условно назвать дискурсивными режимами, поскольку они формируются последовательностями речевых актов. **Дискурсивный режим** — это такой способ использования речевых высказываний, который предполагает регулярную модификацию значений речевых форм, входящих в соответствующий фрагмент, по сравнению с некоторым стандартом. Так, дискурсивный режим иронии предполагает понимание речевых форм, противоположное буквальному значению. Сфера идеологизированного употребления языка предполагает ориентацию на систему

ценностей, объединяющую носителей соответствующей идеологии, что опять-таки влечет регулярное изменение оценочной семантики таких слов и выражений, как *революция, диктатура пролетариата, социальный оптимизм, рабочий, обыватель* и пр.

Заданные теоретические рамки оценки как речевого акта и процедуры оценивания позволяют указать некоторые типы дискурсивных режимов, в которых оценка может как ослабляться, так и усиливаться.

Выявление оценочных характеристик особенно существенно для лингвистической экспертизы, поскольку это необходимо при проведении экспертиз по делам о защите чести, достоинства и деловой репутации, а также по делам об оскорблении и экстремизму. Ограничимся далее по большей части отрицательной оценкой, поскольку она в наибольшей степени существенна для лингвистической экспертизы текста.

3. Дискурсивный режим иронии

Как известно, ирония — это в самом общем смысле «троп, состоящий в употреблении слова в смысле обратном буквальному с целью тонкой или скрытой насмешки» (Ахманова 1969). В ряде случаев иронический контекст модифицирует оценочную составляющую высказывания. В одном из блогов была сделана следующая запись:

> *Я призываю всех к массовым беспорядкам на Лубянке 15го!/ хоть бы/ хоть бы/ хоть бы/ ну погреться же надо будет!!!/ PS а в итоге потом меня обвинит СК в призыве))) ГЫ ГЫ ГЫ.* (14 дек в 9:08)

Слово *беспорядки* имеет в лексическом значении оценочную составляющую[1]:

> **БЕСПОРЯДОК <…> 1.1. *зд. мн.*** Массовое проявление протеста, к-рое сопровождается нарушением общественного порядка, уличными столкновениями и т. п. (БУСРЯ)

В приведенном толковании компонент ‘нарушение общественного порядка, уличные столкновения’ в явном виде указывает на отрицательную оценку. В контексте блога — особенно с учетом современной «охранительной» внутренней политики — слово *беспорядки* носит отчетливый негативный характер[2]. Ср. характерный пример из Паустовского, приводимый в МАСе: *В Таганроге были беспорядки, толпа голодных женщин с детьми разгромила пекарни и продовольственные магазины, и казаки отказались стрелять по толпе*. Разгром пекарен и продовольственных магазинов нельзя рассматривать как что-то положительное.

---

[1] Здесь и далее словарная информация дается в сокращении.

[2] Совсем недавно, слово *беспорядки* выражало положительную оценку: «народные волнения, являвшиеся выражением протеста против власти эксплуататоров» [МАС].

В блоговой коммуникации фраза *Я призываю всех к массовым беспорядкам* не вполне естественна по следующим причинам. Во-первых, она содержит цитацию юридической терминологии. Действительно, юридический термин «призыв к массовым беспорядкам» относится к письменной речи (Ср. фразы, не вполне корректные в обыденной коммуникации: *я призываю возбудить вражду к группе лиц, выделяемой по религиозному признаку; я призываю похитить чужое имущество*).

Во-вторых, призыв к действиям, которые сам говорящий оценивает, скорее, отрицательно, прагматически аномален из-за феномена «иллокутивного самоубийства» по З. Вендлеру (Вендлер 1985). Ср. призывы, некорректные по той же причине: *Я призываю к ограблению магазина!; Я призываю к обману покупателей!* при норме *Я призываю к борьбе за светлое будущее человечества!*

В-третьих, в рассматриваемой фразе цитация юридической терминологии указывает на то, что автор обращает внимание на возможные юридические санкции, которые могут для него последовать после данного акта речевого поведения («метатекстовое» употребление). На метатекстовый характер примера указывает и последующий комментарий блогера: PS *а в итоге потом меня обвинит СК в призыве))) ГЫ ГЫ ГЫ*.

В-четвертых, в исследуемой записи представлена такая мотивация «призыва к массовым беспорядкам» (*хоть бы/хоть бы/хоть бы/ну погреться же надо будет!!!*), которую нельзя рассматривать как серьезное обоснование такой деятельности.

Выявленные характеристики обсуждаемой фразы — использование юридической терминологии в блоговой коммуникации, призыв к действиям, которые сам автор оценивает, скорее, отрицательно, метатекстовый характер употребления, несерьезность обоснования призыва к беспорядкам — все это указывает на то, что в рассматриваемой записи в блоге реализуется семантика самоиронии. Ирония существенно снижает силу негативной оценки, указывая на «несерьезность», нарочитый характер речевого поведения.

4. Дискурсивный режим языковой игры

Как известно, игра снимает многие социальные табу. В литературе неоднократно отмечалось, что во время карнавала снимаются многие культурные и религиозные запреты, поскольку по своему содержанию и функции игра противопоставлена реальности и выводит за рамки этических оценок многие формы обычно табуированного поведения (ср. по этому поводу исследование карнавала как феномена культуры в работах М. М. Бахтина — Бахтин 1965).

Контекст языковой игры ослабляет, а в некоторых случаях и вовсе нивелирует оценочную составляющую многих языковых форм. Так, оценочные словосочетания *дерьмократы, либерасты, толерасты* передают негативную оценку в существенно более ослабленной форме, чем соответствующие полные неигровые предикации: *Демократы — дерьмо, Либералы — пидорасы, Толерантные люди — пидорасы*.

Следует отметить, что игровой режим далеко не всегда снижает негативную оценку. Так, рифмование, использование аллитераций часто никак не затрагивает оценочность контекста, а в некоторых случаях даже ее усиливает.

Например, в листовке правоэкстремистского толка, направленной против мигрантов, во фразе *поймаешь мразь — по морде — хрясь*, сопровождающей соответствующий рисунок, слово с очевидной негативной оценкой *мразь* не меняет своих характеристик в сторону снижения отрицательной оценки объекта оценивания. Причем рифмование *мразь — хрясь*, будучи чисто игровым приемом, в данном случае, скорее, усиливает негативную оценку, чем преуменьшает. Это связано с тем, что слово *хрясь*, указывая на интенсивность физического воздействия (удара) и даже некую удовлетворенность от произведенного действия, передает общую положительную оценку автором листовки (а также ее демонстратором) ситуации избиения неугодного лица. Негативная оценка избиваемого в слове *мразь* поддерживается и усиливается общей положительной оценкой физического воздействия на неугодных, типичной для речевых практик правоэкстремистского дискурса. Ослабление оценочности в игровых контекстах экстремистского дискурса нельзя рассматривать как универсальное правило.

5. Дискурсивный режим общепринятой номинации

В названиях довольно часто присутствует оценочная лексика, однако в общепринятой номинации (в ее внутренней форме) оценочные смыслы ослабляются или теряют свою силу. Оценочная составляющая стандартных номинаций проявляется только как фактор скрытого (имплицитного) воздействия. Например, в официальной номинации «Государственный университет высшая школа экономики» почти все слова — за исключением слова *экономика* — имеют положительные коннотации прагматического характера: прилагательное *государственный* в рассматриваемом контексте связывается с идеей стабильности и основательности (ср. *мой ребенок поступил в государственный вуз, а не в частный*), слово *университет* — с всеобъемлющим знанием и универсальностью обучения, что также воспринимается как что-то положительное; словосочетание *высшая школа* указывает на максимально возможную степень образования и обучения. Понятно, что данное название сконструировано совершенно целенаправленно: как прием имплицитного наведения положительной оценки. Интересно, что языковая практика пошла по неожиданному пути, выбрав лишь одно слово — *Вышка*, — коннотации которого не бесспорно положительные (ср. жаргонизм *вышка* как обозначение «высшей меры наказания» — расстрела).

То же верно и в отношении негативных оценок: негативные оценки в составе стандартной номинации часто нивелируются. Так, в (Успенский 2000: 31) разбирается пример использования стандартной номинации *собака Калин [царь]* («постоянного эпитета») в русском фольклоре. Понятно, что в такой номинации фиксируется идеологическая (и политическая) позиция говорящего:

Говорил ***собака Калин царь*** *да то таковы слова:*
— Ай же старыя казак да Илья Муромец! <…>
— Не служи-тко ты князю Владимиру.
— Да служи-тко ты ***собаке Царю Калину****.*

В последнем случае самому царю Калину приписывается номинация, которая очевидно не могла звучать из его уст. Воздействующий эффект в передаче негативной оценки в таких случаях при постоянном воспроизводстве существенно уменьшается. Это связано с тем, что цель использования такой идеологической номинации состоит в самоотождествлении говорящего с определенной референтной группой, использующей эту номинацию, а не в том, чтобы выразить негативную оценку.

6. Дискурсивный режим неопределенной и определенной референции

Денотативный статус имени (именной группы) как усиливает, так и снижает негативную оценку, которая к нему относится. Отнесенность негативной оценки к конкретной сущности, обозначенной определенной именной группой, усиливает негативную оценку. Так, фраза *Петров — идиот* передает сильный вариант негативной оценки. В то же время отнесенность оценки к неопределенному множеству лиц негативную оценку существенно снижает: *Это сделали какие-то идиоты*. Невозможность ясной атрибуции негативной оценки снижает ее конфликтность, а конкретная референция существенно повышает. Именно поэтому, в практике судебных дел по защите чести и достоинства (ст. 152 ГК РФ) фразы типа *Гаишники берут взятки*, *В министерстве все уже проворовались*, *Руководство компании погрязло в скандалах и склоках* при условии отсутствия политической составляющей не имеют серьезной судебной перспективы. Действительно, конкретные лица не указаны, негативная оценка относится либо к неопределенной именной группе (*гаишники*), либо к именным группам, обозначающим множества людей (*министерство, руководство компании*). Тем самым, конкретные лица опять-таки не названы.

Совершенно иная ситуация с контекстами, в которых референция имени определенна:

> *Начальник второго управления ГИБДД П. Колпаков известный взяточник; Замминистра Егоров был уличен в воровстве бюджетных средств; Скандал был инициирован, как всегда, заместителем Председателя наблюдательного совета компании Давидовичем.*

Снижают негативную оценку и слабоопределенные именные группы (референт известен говорящему, но неизвестен слушающему). Так, фраза *Некоторые компании действовали не вполне законно и сумели получить подряды без тендера* существенно менее конфликтна, чем *Компании Парк-трест и Велосити действовали не вполне законно и сумели получить подряды без тендера*. Разумеется, конфликтность и степень негативности оценки не одно и то же, однако в общем случае снижение конфликтности контекста влечет и уменьшение «негативности» оценки.

***

Дискурсивные режимы, будучи важной характеристикой общения на естественном языке, представляют собой интегральные феномены, сформированные кластерами более универсальных свойств коммуникации, которые воспроизводятся вместе в соответствии с правилами построения, формирования и функционирования того или иного дискурса. К ним следует отнести многие важные оппозиции, уже обсуждавшиеся в рамках теории коммуникативной организации высказывания и текста, в сфере прагматики и логической традиции: «истинное — ложное», «реальность — фиктивность», «настоящее — кажущееся», «неопределенность — определенность», «фон — фигура», «контраст — отсутствие контраста» и др. Так, для дискурсивного режима общепринятой номинации релевантными оказываются две последних оппозиции, поскольку такие номинации являются фоновыми и они в общем случае не становятся фокусом контраста. Для дискурсивного режима иронии существенной оказывается первая оппозиция, поскольку ирония предполагает, что говорящий имеет в виду ровно противоположное. Дискурсивный режим языковой игры позволяет противопоставить реальное, настоящее и фиктивное, кажущееся. Неопределенная референция связана с категорий неопределенности и часто реализуется как фон в отсутствии контраста.

Указанные оппозиции могут использоваться как диагностические для выявления дискурсивного режима: так, фокусирование актанта ситуации или, наоборот, перемещение его в фоновый компонент сообщения может свидетельствовать о смене дискурсивного режима. Противоречия в авторской оценке типичны для иронии. Именно в этом направлении можно видеть возможный алгоритм выявления дискурсивного режима для программ автоматической обработки текста. Введение модуля выявления дискурсивного режима в перспективе могло бы существенно уточнить результаты автоматического мониторинга тональности дискурса, основанного на словарной информации.

Рассмотренные дискурсивные режимы не исчерпывают весь список таких феноменов. Так, на статус дискурсивного режима могут претендовать цитация, различные формы реферирования содержания текста, метатекстовые употребления, сфера идеологизированного функционирования речевых форм. Влияние дискурсивного режима испытывает не только оценка, но и многие другие семантические категории — как речевые акты, так и различные способы «упаковки» многомерного смысла высказывания и текста.

Литература

1. *Апресян Ю. Д.* Коннотация как часть прагматики слова // Апресян Ю. Д. Интегральное описание языка и системная лексикография. Т. II. М., 1995.
2. *Ахманова О. С.* Словарь лингвистических терминов. М., 1969.
3. *Баранов А. Н.* Аксиологические стратегии в структуре языка (паремиология и лексика) // Вопросы языкознания, 1989, № 3.

4. *Бахтин М. М.* Творчество Франсуа Рабле и народная культура средневековья и Ренессанса. М., 1965.
5. *БУСРЯ* — Большой универсальный словарь русского языка / Под ред. В. В. Морковкина. М., 2016.
6. *Вендлер З.* Иллокутивное самоубийство // Новое в зарубежной лингвистике: Вып. 16. Лингвистическая прагматика. М.: Прогресс, 1985.
7. *Вольф Е. М.* Функциональная семантика оценки. М., 1985. [Переиздание: Вольф Е. М. Функциональная семантика оценки. 2 изд., испр. и доп. М., 2002.]
8. *Шмелев Д. Н.* Слово и образ. М., 1964.
9. *МАС* — Словарь русского языка в 4-х тт. / под ред. Евгеньевой А. П. М., 1985–1988 [«Малый академический словарь»]
10. *Серль Дж.* Классификация иллокутивных актов // Новое в зарубежной лингвистике. Вып. XVII. Теория речевых актов. М., 1986.
11. *Телия В. Н.* Экспрессивность как проявление субъективного фактора в языке и ее прагматическая ориентация // Языковые механизмы экспрессивности. М., 1991.
12. *Успенский Б. А.* Поэтика композиции. Санкт-Петербург, 2000.
13. *Хэар Р. М.* Дескрипция и оценка // Новое в зарубежной лингвистике. Вып. XVI. М., 1985.

References

1. *Akhmanova O. S.* (1969), Dictionary of linguistic terms [Slovar' lingvisticheskikh terminov], Moscow.
2. *Apresjan Ju. D.* (1995), Connotation as a part of word pragmatics [Konnotacija kak chast' pragmatiki slova], Integrative description of language and system lexicography [Integral'noe opisanie jazyka i sistemnaja leksikograija], Moscow.
3. *Bakhtin M. M.* (1965), Works of François Rabelais and folk culture of middle ages and the Renaissance [Tvorchestvo Fransua Rable i narodnaja kul'tura srednevekovja i renesansa], Moscow.
4. *Baranov A. N.* (1989), Evaluation strategies in language (proverbs and lexicon) [Aksiologicheskie strategii v strukture jazyka (paremiologija i leksika)], Problems of linguistics [Voprosy jazykoznanija], № 3, pp. 74–90.
5. *BUSRJA* — Large universal dictionary of Russian ed. by V. V. Morkovkin [Bol'shoj universal'nyj tolkovyj slovar' russkogo jazyka pod red. V. V. Morkovkina] (2016), Moscow.
6. *Hare R. M.* (1985), Description and evaluation [Deskripcija i ocenka], New in foreign linguistics [Novoe v zarubezhnoj lingvistike], Vol. XVI. Moscow, pp. 183–195.
7. *MAS* — Dictionary of Russian Language in 4 volumes [Slovar' russkogo jazyka v 4-kh tomakh] (1985–1988), Russian Language, Moscow.
8. *Searle J.* (1986), A classification of illocutionary acts [Klassifikacija illokutivnykh aktov], New in foreign linguistics [Novoe v zarubezhnoj lingvistike], Vol. XVII, Moscow, 1986, pp. 170–194.
9. *Shmelev D. N.* (1964), Word and image [Slovo i obraz], Moscow.

10. *Teliya V. N.* (1991), Expressivity as a reflection of subjective factor in language and its pragmatic orientation [Ekspessivnost' kak proyavleniye subjektivnogo faktora v jazyke], Language mechanisms of expressivity [Jazykovyje mekhanizmy ekspressivnosti], Moscow.
11. *Uspenskij B. A.* (2000), Poetics of composition [Poetika kompozicii], Saint-Petersburg.
12. *Vendler Z.* (1985), Illocutionary suicide [Illokutivnoe samoubujstvo], New in foreign linguistics [Novoe v zarubezhnoj lingvistike], Vol. 16, Moscow, Progress.
13. *Volf E. M.* (1985), Functional semantics of evaluation [Funkcional'naja semantika ocenki], Moscow.

VERY LARGE RUSSIAN CORPORA: NEW OPPORTUNITIES AND NEW CHALLENGES

Benko V. (vladob@juls.savba.sk)

Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics, Bratislava, Slovakia

Zakharov V. P. (v.zakharov@spbu.ru)

St. Petersburg State University;
Institute for Linguistic Studies, RAS, St. Petersburg, Russia

Our paper deals with the rapidly developing area of corpus linguistics referred to as *Web as Corpus* (*WaC*), i.e., creation of very large corpora composed of texts downloaded from the web. Some problems of compilation and usage of such corpora are addressed, most notably the "language quality" of web texts and the inadequate balance of web corpora, with the latter being an obstacle both for corpus creators, and its users. We introduce the *Aranea* family of web corpora, describe the various processing procedures used during its compilation, and present an attempt to increase the size of its Russian component by the order of magnitude. We also compare its contents from the user's perspective among the various sizes of the Russian *Aranea*, as well as with the other large Russian corpora (*RNC*, *ruTenTen* and *GICR*). We also intent to demonstrate the advantage of a very large corpus in linguistic analysis of low-frequency language phenomena in linguistics, such as usage of idioms and other types of fixed expressions.

Keywords: web corpora, *WaC* technology, representativeness, balance, evaluation

СВЕРХБОЛЬШИЕ КОРПУСЫ РУССКОГО ЯЗЫКА: НОВЫЕ ВОЗМОЖНОСТИ И НОВЫЕ ПРОБЛЕМЫ

Бенко В. (vladob@juls.savba.sk)

Словацкая академия наук, Институт языкознания им. Людовита Штура, Братислава, Словакия

Захаров В. П. (v.zakharov@spbu.ru

Санкт-Петербургский государственный университет; Институт лингвистических исследований РАН, Санкт-Петербург, Россия

В статье обсуждается одно из активно развиваемых направлений в корпусной лингвистике — создание корпусов большого объема на основе текстов из веба. Показаны их возможности в исследовании и описании устойчивых сочетаний. Описываются технология и проблемы их создания. Обсуждаются проблемы таких корпусов, которые ставят вопросы как перед разработчиками корпусов, так и перед пользователями, а именно, проблемы морфологической разметки и сбалансированности корпусов.

Ключевые слова: веб-корпусы, *WaC* технология, репрезентативность, сбалансированность, оценка

0. Introduction

Quantitative assessment of language data has always been an area of great interest for linguists. And not only for them: as early as in 1913, the Russian mathematician A. A. Markov counted the frequencies of letters and their combinations in the Pushkin's *Eugene Onegin* novel, and calculated the lexical probabilities in the Russian language [Markov, 1913]. With the advent of first computers, the usage of quantitative methods in linguistic research has accelerated dramatically [Piotrovskiy 1968; Golovin 1970; Alekseev 1980; Arapov 1988], aiding in creation of frequency dictionaries[1] and in other research activities of theoretical and applied nature [Frumkina 1964, 1973].

The next step in using quantitative methods in language research has been done within an area of corpus linguistics. The results of corpus queries are usually accompanied by the respective statistical information. Advanced corpus management systems provide for obtaining all sorts of statistical data, including those of linguistic categories and metadata. Combination of quantitative methods, distributional analysis and contrastive studies is becoming the basis of new corpus systems that could be referred to as "intellectual". Their functionalities include automatic extraction of collocations, terms, named entities, lexico-semantic groups, etc. In fact, corpus linguistics based on formal language models and quantitative methods is "learning" to solve intellectual semantic tasks.

Assuming that one of the main features of a representative corpus is its size, then a 100-million token corpus, considered a standard at the beginning of this century, now appears in many cases to be insufficient to receive relevant statistical data. To study and adequately describe multi-word expressions consisting of medium or low-frequency words, it is necessary to apply large and even very large corpora. In the context of this paper, we call a corpus "very large" if its size exceeds 10 billion tokens[2].

---

[1] It should be noted, however, that first frequency dictionaries have been compiled well in the pre-computer era, in the end of the19th century [Kaeding 1897].

[2] In Russian, we suggest the term "сверхбольшой корпус".

1. Web as Corpus

Nowadays, the "big data" paradigm became very popular. According to Wikipedia, "*Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate*"[3]. This "big data" now seem to have approached the corpus linguistics.

Compilation of traditional corpora is usually a laborious and rather slow process. As soon as the need for larger corpora has been recognized, it became clear that the requirements of the linguistic community cannot be easily satisfied by the traditional resources of corpus linguistics. This is why many linguists in the process of their research turned to Internet search services. But using search engines as corpus query systems is associated with many problems (cf. [Kilgarriff 2007; Belikov et al. 2012])—this is where the idea of *Web as Corpus* (*WaC*), i.e., creation of language corpora based on the web-derived data has been born. It was apparently for the first time explicitly articulated by Adam Kilgarriff [Kilgarriff 2001; Kilgarriff, Grefenstette 2003].

In early 2000s, a community called *WaCky!*[4] was established by a group of linguists and IT specialists who were developing tools for creation of large-scale web corpora. During the period of 2006–2009, several *WaC* corpora were created and published, including the full documentation of the respective technology, with each containing 1–2 billion tokens (*deWaC*, *frWaC*, *itWaC*, *ukWaC*) [Baroni et. al 2009].

In 2011, the *COW*[5] (*COrpora from the Web*) project started at the Freie Universität in Berlin. Within its framework, English, German, French, Dutch, Spanish and Swedish corpora have been created. In the 2014 edition (*COW14*) of the family, sizes of some corpora reached almost 10 billion tokens, while the German corpus has 20 billion tokens [Schäfer, Bildhauer 2012; Schäfer 2015]. These corpora are accessible (for research purposes) via the project web portal[6]. The site also provides English, German, Spanish and Swedish corpus-based frequency lists.

Large number of *WaC* corpora has been created and/or made available within the framework of the *CLARIN* Project in Slovenia (Jožef Stefan Institute). Besides the respective South Slavic languages (*bsWaC*, *hrWaC*, *slWaC*, *srWaC*) [Ljubešić, Erjavec 2011; Ljubešić, Klubička 2014], corpora for many other languages, including Japanese, are available there. Their sizes vary between 400 million and 2 billion tokens. Most of the corpora are accessible[7] under *NoSketch Engine*[8] without any restrictions.

None of the projects mentioned, however, includes the Russian language.

The largest number of *WaC* corpora was created by Lexical Computing Ltd. (Brighton, UK & Brno, Czech Republic) company that made them available within

---

[3] https://en.wikipedia.org/wiki/Big_data

[4] http://wacky.sslmit.unibo.it/

[5] http://hpsg.fu-berlin.de/cow/

[6] https://webcorpora.org/

[7] http://nl.ijs.si/noske/index-en.html

[8] https://nlp.fi.muni.cz/trac/noske

Sketch Engine[9] environment. At the time of writing this paper (April 2016), these corpora covered almost 40 languages, including Russian, and their sizes varied between 2 million and 20 billion tokens. The size of the largest Russian *ruTenTen* corpus was 18.3 billion tokens [Jakubíček et al. 2013].

From today's perspective, we can see that the *WaC* technology has succeeded. Related set of application programs that represent effective implementation of this technology has been published, including tools for web crawling, data cleaning and deduplication, with many of them under free or open-source licenses (*FLOSS*) that made the technology available also for underfunded research and educational institutions in Central and Eastern Europe.

There are, however, also other approaches to creation of very large corpora. One of them—based on massive digitization of books from public libraries—has been attempted by Google (available via *Google Books Ngram Viewer*[10]) [Zakharov, Masevich 2014]. Another possibility is creating corpora based on the integral web collections, such as the *General Internet Corpus of Russian*[11] (*GICR, 19.7 billion tokens*) [Belikov et al. 2013], that is composed of blogs, social media, and news.

2. WaC "How To"

To create a web corpus, we usually have to perform (in a certain sequence) operations as follows:

- Downloading large amounts of data from the Internet, extracting the textual information, normalizing encoding
- Identification the language of the downloaded texts, removing the "incorrect" documents
- Segmenting the text into paragraphs and sentences
- Removing duplicate content (identical or partially identical text segments)
- Tokenization—segmenting the text into words
- Linguistic (morphological, and possibly also syntactic) annotation—lemmatization and tagging
- Uploading the resulting corpus into the corpus manager (i.e., generating the respective index structures) that will make the corpus accessible for the users.

With the exception of first two, all other operations have been already included (to a certain extent) in the process of building traditional corpora. It is therefore often possible to use existing tools and methodology of corpus linguistics, most notably for morphological and syntactic annotation.

Downloading data from the web is usually performed by one of two standard methodologies that differ in the way how the URL addresses of the web pages to be downloaded are retrieved.

---

[9] http://www.sketchengine.co.uk

[10] https://books.google.com/ngrams

[11] http://www.webcorpora.ru/en

(1) Within the method described in [Sharoff 2006], a list of medium-frequency words is used to generate random n-tuples that are subsequently iteratively submitted to a search engine. Top URL addresses delivered within each search are then used to download the data for the corpus. The process can be partially automated by the *BootCaT*[12] program [Baroni, Bernardini 2004].

(2) The second method is based on scanning ("crawling") the web space by means of a special program—crawler—that uses an initial list of web addresses provided by the user and iteratively looks for new URLs by analysing the hyperlinks at the already downloaded web pages. The program usually works autonomously and may also perform encoding/language identification and/or deduplication on the fly, which makes the whole process very efficient and allows in a relatively short time (several hours or days) download textual data containing several hundreds of millions tokens. Two most popular programs used for crawling the web corpora are the general-purpose *Heritrix*[13] and a specialized "linguistic" crawler *SpiderLing*[14] [Suchomel, Pomikálek 2012].

Each of the methods mentioned above has its pros and cons, with the former being more suitable for creation of smaller corpora (especially if the corpus is geared towards a specific domain), while the latter is usually used to create very large corpora of several billions of tokens in size.

3. The *Aranea* Web Corpora Project: Basic Characteristics and Current State

The *Aranea*[15] family presently consists of (comparable) web corpora created by the *WaC* technology for 14 languages in two basic sizes. The *Maius* ("larger") series corpora contain 1.2 billion tokens, i.e. approximately 1 billion words (tokens starting with alphabetic characters). Each *Minus* ("smaller") corpus represents a 10% random sample of the respective *Maius* corpus. For some languages, region-specific variants also exist that, e.g., increase the total number of Russian corpora to six. *Araneum Russicum Maius & Minus* include Russian texts downloaded from any internet domains, *Araneum Russicum Russicum Maius & Minus* contain only texts extracted from the *.ru* and *.рф* domains, and *Araneum Russicum Externum Maius & Minus* are based on texts from "non-Russian" domains, such as *.ua*, *.by*, *.kz*, etc. For more details about the *Aranea* Project see [Benko 2014].

According to our experience, a Gigaword corpus can be created by means of *FLOSS* tools in a relatively short time, even on a not very powerful computer. After the processing pipeline had been standardized, we were able to create, annotate and publish

[12] http://bootcat.sslmit.unibo.it/

[13] https://webarchive.jira.com/wiki/display/Heritrix

[14] http://corpus.tools/wiki/SpiderLing

[15] http://ella.juls.savba.sk/aranea_about

a corpus for a new language in some 2 weeks (provided that the respective tagger was available).

The situation, however, has changed when we wanted to increase the corpus size radically. We decided to create a corpus of a *Maximum* class, i.e., "as much as can get". Our attempt to create the Slovak and Czech *Maximum* corpora revealed that the limiting factor was the availability of the sufficient amounts of texts for the respective languages in Internet. With standard settings for *SpiderLing* and after several months of crawling, we were able to gather only some 3 Gigawords for Slovak and approximately 5 Gigawords for Czech.

To verify the feasibility of building very large corpora within our computing environment, we decided to create *Araneum Maximum* for a language, where sufficient amount of textual data in Internet is expected. The Russian language has been chosen for this experiment, and the lower size limit was set to 12 billion tokens, i.e., ten times the size of the respective *Maius* corpus.

It has to be noted that the work was not to be started from scratch, as the data of existing Russian *Aranea* had been utilized. After joining all available Russian texts and deduplicating them at the document level, we received approximately 6 billion tokens, i.e., seemingly half of the target corpus size. It was, however, less than that, as the data had not been dedulicated at the paragraph level yet.

The new data was crawled by the (at that time) newest version 0.81 of *SpiderLing*, and the seed URLs were harvested by *BootCaT* as follows:

(1) A list of 1,000 most frequent adverbs extracted from the existing Russian corpus was sorted in random order (adverbs have been chosen as they do not have many inflected forms and usually have rather general meaning).

(2) For each *BootCat* session, 20 adverbs were selected to generate 200 *Bing* queries (three adverbs in each), and requesting to get the maximal amount of 50 URLs from each query. This procedure has been repeated five times, totalling in 1,000 *Bing* queries.

The number of URLs harvested by a single *BootCaT* session in this way was usually close to the theoretical maximum of 50,000, but it decreased to some 40,000–45,000 after filtration and deduplicaction. The resulting list was sorted in random order and iteratively used as seed for *SpiderLing*.

To create a *Maius* series corpus, we always tried to gather approximately 2 billion tokens of data, so that the target 1.2 billion can be safely achieved after filtration and deduplication. For "large" languages, this could be reached during first two or three days of crawling. As it turned out later, we were quite lucky not to reach the configuration limits of our server, most notably the size of RAM (16 GB). As all data structures of *SpiderLing* are kept in main memory, when trying to prolong the crawling time for the Russian the memory limit has been reached only after approximately 80–90 hours of crawling. Though some memory savings tricks are described in the *SpiderLing* documentation, we, nonetheless, had to opt for a "brute force" method by restarting the crawling several times from scratch, knowing that lots of duplicate data would be obtained.

In total, 12 such crawling iterations (with some of them consisting of multiple sessions) have been performed, during which we experimented with the number of seed URLs ranging from 1,000 to 40,000.

To speed up the overall process, another available computer was used for cleaning, tokenization, partial deduplication and tagging of the already downloaded lots of data. Moreover, the most computationally-intensive operations (tokenization and tagging) have been performed in parallel, taking the advantage of the multiple-core processor of our computer. The final deduplication has been performed only after all data has been joined into one corpus.

Our standard processing pipeline contains the steps described in Tables 1 and 2.

Table 1. Processing of a typical new lot (one of 12)

Operation	Output	Processing time (hh:mm)
Data crawling by *SpiderLing* (2 parallel processes) with integrated boilerplate removal by *jusText*[16] [Pomikálek 2011] and identification of exact duplicates	2,958,522 docs 39.68 GB	cca 86 hours
Deleting duplicate documents identified by *SpiderLing*	2,058,810 docs 18.15 GB	0:27
Removing the survived HTML markup and normalization of encoding (Unicode spaces, composite accents, soft hyphens, etc.)		0:30
Removing documents with misinterpreted utf-8 encoding	2,054,827 docs	0:41
Tokenization by *Unitok*[17] [Michelfeit et al. 2014] (4 parallel processes, custom Russian parameter file)	1,611,313,889 tokens 19.88 GB	4:04
Segmenting to sentences (rudimentary rule-based algorithm)		0:29
Deduplication of partially identical documents by *Onion*[18] [Pomikálek 2011] (5-grams, similarity threshold 0.9)	1,554,837 docs 1,288,238,029 tokens (20.05% removed) 17.23 GB	1:23

[16] http://corpus.tools/wiki/Justext

[17] http://corpus.tools/wiki/Unitok

[18] http://corpus.tools/wiki/Onion

[19] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[20] http://corpus.leeds.ac.uk/mocky/

[21] http://nl.ijs.si/ME/V4/msd/html/msd-ru.html

[22] http://ella.juls.savba.sk/aranea_about/aut.html

Operation	Output	Processing time (hh:mm)
Conversion all utf-8 punctuation characters to ASCII and changing all occurrences of "*ë*" to "*e*" (to make the input more compatible with the language model used by the tagger).		0:53
Tagging by *Tree Tagger*[19] [Schmid 1994] with language model trained by S. Sharoff[20] (4 parallel processes)	39.06 GB	8:26
Recovering the original utf-8 punctuation and "ë" characters		0:53
Marking the out-of-vocabulary (*OOV*) tokens (*ztag*)	82,786,567 tokens marked *OOV* (6.43%)	1:09
Mapping the "native" *MTE*[21] tagset to "PoS-only" *AUT*[22] tagset	46.39 GB	1:09

Table 2. Final processing

	Output	Processing time (hh:mm)
Joining all parts of data (old data + 12 new lots, some of them accessed via Ethernet at a different machine)	37,956,781 docs 26,720,417,271 tokens 932.80 GB	10:42
Deduplication of partially identical documents by *Onion* (5-grams, similarity threshold 0.9)	24,509,170 docs 17,322,616,899 tokens (35.17% removed) 602.33 GB	19:12
Deduplication of partially identical paragraphs by *Onion* (5-grams, similarity threshold 0.9)	13,704,863,990 tokens (20.88% removed) 482.04 GB	27:07
Compilation by *NoSketch Engine*	249.78 GB of index structures	79:54

4. Experimenting with the New Corpus

At the end of all the processing mentioned, we indeed succeeded to create a very large Russian corpus of the expected size—its characteristics (as displayed by *NoSketch Engine*) are shown in Fig. 1.

Corpus *Araneum Russicum Maximum (Russian, 16.04) 13,7 G* – statistics and info

Russian Web (crawled 2013 to 2016, version 1.3.70) 13,7 G (build #a068)

Counts	
Tokens	13,704,863,994
Words	10,945,698,722
Sentences	798,912,811
Paragraphs	299,974,845
Documents	24,509,166

General info	
Corpus description	Document
Language	Russian
Encoding	UTF-8
Compiled	04/23/2016 15:56:17
Tagset	Description

Lexicon sizes	
word	43,132,672
lemma	40,074,411
atag	12
tag	1,140
ztag	2
lc	37,345,635
lemma_lc	34,686,361

Fig. 1. New Corpus Info

Within the context of *NoSketch Engine*, a token is considered "word" if it begins with an alphabetic character (in any script recognized by Unicode). It must be also noted that the lemma lexicon contains large proportion of out-of-vocabulary items that could not have been lemmatized.

In the following text, we will demonstrate the usefulness of a very large corpus for studying rare language phenomena, such as phraseology.

4.1. Chasing Fixed Expressions

In small corpora, many idioms often appear—if ever—in singular ("hapax") occurrences that make it difficult to draw any relevant linguistic conclusions. Moreover, idioms and other fixed expressions are often subject to lexical and/or syntactic variation, where the individual members of the expressions change within a fixed syntactic formula, or the same set of lexical units create different syntactic structures [Moon 1998]. It is most likely without exaggeration to claim that idioms having lexical and syntactic variants represent the majority of cases. Lots of (Russian) examples can be shown: *беречь/хранить как зеницу ока; беречь пуще глаза; мерить одной мерой/меркой, мерить на одну меру/мерку; ест за троих, есть в три горла; драть/сдирать/содрать шкуру (три/две шкуры), драть/сдирать/содрать по три (две) шкуры; хоть в землю зарройся, хоть из-под земли достань; брать/взять (забирать/забрать) в [свои] руки, прибирать/прибрать к рукам; сталкивать/столкнуться лицом к лицу, носом к носу, нос в нос, лоб в лоб.*

The description of variant multi-word expressions in dictionaries is naturally much less complete in comparison with fixed phrasemes. And, only large and very large corpora can help us to analyse and describe this sort of variability in full.

Now we shall try to demonstrate the possibilities given by *Araneum Russicum Maximum* on three examples. Let us take fixed expressions described in dictionaries and show how they behave in various corpora.

4.2. *"Щёки как у хомяка"*[23]

The *Russian National Corpus* (*RNC*[24], 265 M tokens[25]) gives 5 occurrences of "*щёки как*": *как у матери, как у бульдога, как у пророка, как у тяжко больного, как у меня*. As it can be seen, all of them are singular occurrences (hapax legomena), and no occurrence of *как у хомяка* has been found.

Let us have a look what can be found in other corpora. While the smaller *Aranea* provide even less information, *Araneum Russicum Maximum* confirms the dictionary data, and *ruTenTen* and *GICR* corpora make it even more convincing. Besides *как у хомяка*, they also add *как у бульдога*, *как у бурундука* and *как у матрешки*, as well as several other (less frequent) comparisons.

Table 3. "*Щёки как у…*"

	щёки/ щеки как у...	**хомяка/ хомячка**	**буль-дога**	**бурун-дука**	**мат-решки**
Araneum Russicum Minus	1	–	1	–	–
Araneum Russicum Maius	1	–	1	–	–
Araneum Russicum Maximum	33	6	1	4	2
ruTenTen	45	24	4	–	1
GICR	126	84	3	5	1

4.3. *"Щёки из-за спины видны"*[26]

RNC gives just one example of *щеки из-за...*: *щеки из-за ушей видны*.
The other corpora give the following:

Table 4. "*Щёки из-за…*"

	щёки/щеки из за...	**спины видны/ видать/торчат**	**ушей видны/ видать/торчат**
Araneum Russicum Minus	–	–	–
Araneum Russicum Maius	6	3	–
Araneum Russicum Maximum	27	7	5
ruTenTen	30	20	6
GICR	65	40	23

[23] "*cheeks like a hamster*"

[24] http://www.ruscorpora.ru/en/search-main.html

[25] This number is not directly comparable with other corpora, as punctuation characters are not considered tokens in *RNC*.

[26] "*cheeks visible from behind*"

The very large corpora not only provide much more evidence, but also add several interesting variants of "*щеки из-за…*": *увидеть можно, просматриваются, вылезают, сияют румянцем; щек из-за спины видно не было*, etc.

4.4. "*Чистой воды…*"[27]

The idiomatic expression *чистой* or *чистейшей воды* is described in the dictionary as "*о ком или чем-либо, полностью соответствующем свойствам, качествам, обозначенным следующим за выражением существительным*" [BED 1998]. But if we want to extract the relevant information on the most frequent noun collocates of this expression from *RNC*, we mostly get 2–3 examples for each noun: *авантюрист, блеф, гипотеза, демагогия, монополизм, мошенничество, популизм, провокация, садизм, спекуляции, фантастика, хлестаковщина,* etc.

What can be observed in larger corpora? When comparing frequency ranks of expressions with different nouns derived from large corpora, we can see that they are more or less similar, while the data received from small corpora can differ significantly. Nouns appearing at the top positions of the ranked frequency lists derived from the large corpora (*выдумка, вымысел, лохотрон, обман, пиар, профанация, развод, спекуляция*) are usually missing in the output from smaller corpora. On the other hand, top words obtained from *Araneum Russicum Minus* (*чудодействие, грабеж, подстава*) are ranked 50, or even 500 in large corpora. We can also see that the total weight of expressions with significant frequencies (4 or more within the framework of our experiment) is greater in large corpora (Table 5).

Table 5. Frequencies of "*чистой/читейшей воды + noun*" expressions in various corpora

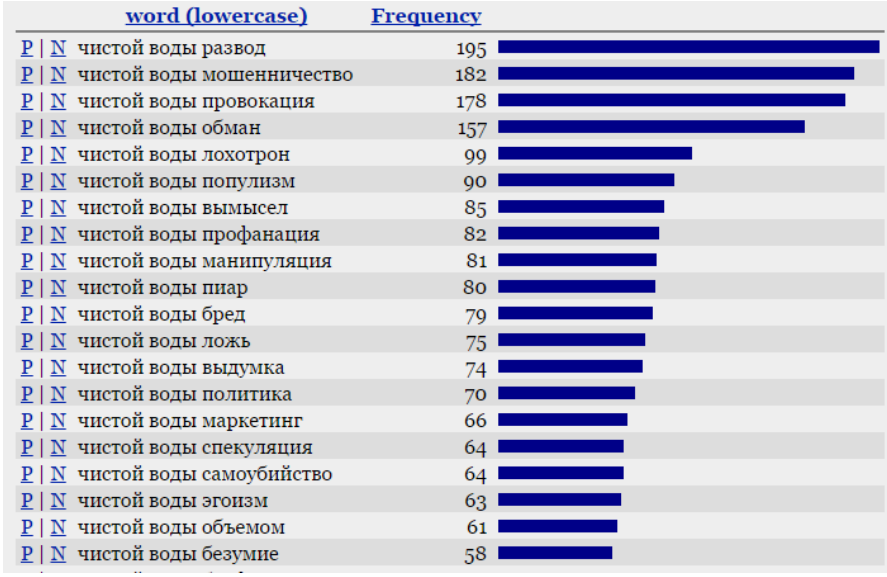
corpus size in tokens	**Araneum Russicum Minus 120 M**	**Araneum Russicum Maius 1.2 G**	**Araneum Russicum Maximum 13.7 G**	**ruTenTen 18.3 G**
total expressions	**146**	1,256	10,441	15,548
unique expressions	26	692	3,264	≥ 5,000[28]
total expressions with f >3	12 (8.2%)	450 (35.8%)	6,841 (65.5%)	9,370 (60.3%)
unique expressions with f >3	2 (7.7%)	54 (7.8%)	449 (13.8%)	668 (13.4%)

[27] "*of the clear water*"

[28] Only first 5,000 items of frequency distributions are shown in Sketch Engine.

The corpus evidence, however, shows that the *чистой воды* expression is also used in its direct meaning. In fact, there are two direct meanings of "*чистой воды*" present there: "*вода чистая, без примесей*", and "*чистая, свободная ото льда или водной растительности*". The interesting fact is, that practically in all cases where *чистой воды* precedes the respective noun, its meaning is idiomatic (Fig. 2).

In *Araneum Russicum Maximum*, out of 449 different analysed expressions with total count of 6,841, less than 10 contained non-idiomatic use of "*чистой воды*" (associated with *объем/температура* or *озеро/море/океан*). And, the majority of the respective nouns have a negative connotation: *абсурд, авантюра, агрессия, алчность, бандит, блеф, богохульство, болтология, бред, брехня, бытовуха, вампиризм, вкусовщина, вранье, глупость, госдеповец, графоманство, демагог, диктатура, жульничество, заказняк, зомбирование, идеализм, извращение, издевательство, инквизиция, кальвинизм, капитализм, кидалово, копипаст, коррупция, лапша, липа, литература, популизм, порнография, пропаганда, развод, расизм, рвач, русофобия, садизм, фарисейство, фарс, фашизм*, etc. Some of them are receiving this negative connotation especially within this expression (*кальвинизм, капитализм, копипаст, лапша, липа, литература, пропаганда* etc.)



word (lowercase)	Frequency
P | N чистой воды развод	195
P | N чистой воды мошенничество	182
P | N чистой воды провокация	178
P | N чистой воды обман	157
P | N чистой воды лохотрон	99
P | N чистой воды популизм	90
P | N чистой воды вымысел	85
P | N чистой воды профанация	82
P | N чистой воды манипуляция	81
P | N чистой воды пиар	80
P | N чистой воды бред	79
P | N чистой воды ложь	75
P | N чистой воды выдумка	74
P | N чистой воды политика	70
P | N чистой воды маркетинг	66
P | N чистой воды спекуляция	64
P | N чистой воды самоубийство	64
P | N чистой воды эгоизм	63
P | N чистой воды объемом	61
P | N чистой воды безумие	58

Fig. 2. Frequency distribution of right-hand noun collocates of *чистой воды* in *Araneum Russicum Maximum*

On the other hand, if *чистой воды* is located after the corresponding noun, the share of its direct meaning is as much as 80% (*литр чистой воды, стакан чистой воды, количество, подача, перекачивание, источник, резервуар, глоток, кран чистой воды*, etc.)

5. Conclusions and Further Work

As it can be seen, very large corpora enable much deeper analysis that is not possible with corpora of smaller size. We can also say that, starting from a certain size of corpora, the results of these studies can be seen as representative. On the other hand, we do not want to state that web corpora could fully replace the traditional ones. They can, however, be really very large and reflect the most "fresh" changes of the language.

Our experiment has also shown that not everything is that simple. The problems encountered can be divided into three parts: problems of linguistic annotation (lemmatization and tagging), problems of metadata (tentatively referred to as "meta-annotation"), and technical problems related to deduplication and cleaning. It is clear that the traditional TEI-compliant meta-annotation cannot be performed in web corpora, as they lack the explicit necessary bibliographic data. In practice, we can get data only with minimal bibliographic annotation in terms of web (domain name, web page publication or crawl date, document size, etc.), and traditional concepts of representativeness and/or balance are hardly applicable. What we can get is the volume, but the question of "quality" remains without an answer. Both the nature of textual data and the imbalance of web corpora make the question of assessing the results of analyses based on such corpora open.

A new methodology based on the research has to be developed yet. We believe that such methods should include both quantitative and qualitative assessments from the perspective of applicability of very large corpora in various types of linguistic research. It might also be useful to compare contents of web corpora with the existing traditional corpora, as well as with frequency dictionaries. It is also necessary to take into account the technical aspects, such as "price vs. quality" relation.

Our experiment aimed to create the Russian *Araneum Maximum* has shown that though some technical problems related to the computing power of our equipment (two quad-core Linux machines witch 16 GB RAM and 2 TB of free disk space each, joined by a Gigabit Ethernet line, and having a 100 Mbit Internet connection), do exist, they could be eventually solved. The bottleneck of the process was the final deduplication by *Onion* that needed 56 GB of RAM, and had to be performed on a borrowed machine. After minor modifications of our processing pipeline, we were able to perform all other operations, including the final corpus compilation by the *NoSketch Engine* corpus manager using our own hardware.

The first results based on our new corpus show that in comparison the *RNC*, *Araneum Russicum Maximum* can provide much more data on rare lexical units and fixed expressions of different kinds and allows for linguistic conclusions. On the other hand, our experience shows that lexis typical for fiction and poetry seems to be underrepresented in our corpus.

Our next work will be targeted both at the increase of the size of our corpus, and also at improving its "quality"—by better filtration, normalization and linguistic annotation. Here we hope to apply methods of crowd-sourcing (e.g., verifying the morphological lexicons by students). The other serious task will be the classification of the texts according to web genres, so that the balance of the corpus could be—at least partially—controlled.

Acknowledgements

This work has been, in part, supported by the Slovak Grant Agency for Science (VEGA Project No. 2/0015/14), and by the Russian Foundation for the Humanities (Project No. 16-04-12019).

References

1. *Alexeev P. M.* (1980), Statistical lexicography [Statisticheskaya lexikografiya], Moscow.
2. *Arapov M. V.* (1988), Quantitative linguistics [Kvantitativnaya lingvistika], Moscow.
3. *Baroni M., Bernardini S.* (2004), BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC 2004.
4. *Baroni M., Bernardini, S., Ferraresi A., Zanchetta E.* (2009), The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. Language Resources and Evaluation 43 (3), pp. 209–226.
5. *BED* (1998), Kuznetsov S. A. (Ed.) Big Explanatory Dictionary of the Russian Language [Bol'shoj tolkovyj slovar' russkogo yazyka], St. Petersburg: Norint.
6. *Belikov V., Selegey V., Sharoff S.* (2012). Preliminary considerations towards developing the General Internet Corpus of Russian // Komp'juternaja lingvistika i intellectual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog–2012» [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2012»]. Moscow, RGGU, pp. 37–49.
7. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.* (2013), Corpus as language: from scalability to register variation, [Korpus kak yazyk: ot masshtabiruyemosti k differentsial'noy polnote], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2013" [Komp'juternaja lingvistika i intellektual'nye tekhnologii: po materialam ezhegodnoy mezhdunarodnoj konferentsii "Dialog 2013"], vol. 12 (19), Moscow, RGGU, pp. 84–95.
8. *Benko V.* (2014), Aranea: Yet Another Family of (Comparable) Web Corpora, In: Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8–12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, pp. 257–264, ISBN: 978-3-319-10815-5.
9. *Frumkina R. M.* (1964), Statistical methods of lexica research [Statisticheskiye metody izucheniya leksiki], Moscow.
10. *Frumkina R. M.* (1973), The role of statistical methods in modern linguistic researches [Rol' statisticheskikh metodov v sovremennykh lingvisticheskikh issledovaniyakh], Moscow.
11. *Golovin B. N.* (1970), Language and statistics [Yazyk i statistika], Moscow.
12. *Jakubíček M., Kilgarriff A., Kovář V., Rychlý P., Suchomel V.* (2013), The TenTen Corpus Family, 7th International Corpus Linguistics Conference, Lancaster, July 2013.
13. *Kaeding F. W.* (1897), Häufigkeitswörterbuch der deutschen Sprache. Steglitz b. Berlin.

14. *Kilgarriff A.* (2001), Web as corpus, in P. Rayson, A. Wilson, T. McEncry, A. Hardic and S. Klioja (eds.) Proceedings of the Corpus Linguistics 2001 Conference, Lancaster (29 March—2 April 2001). Lancaster: UCREL, pp. 342–344.
15. *Kilgarriff A., Grefenstette G.* (2003), Introduction to the Special Issue on Web as Corpus. Computational Linguistics 29 (3), 2003. Reprinted in Practical Lexicography: a Reader. Fontenelle, T. (Ed.) Oxford University Press. 2008.
16. *Kilgarriff A.* (2007), Googleology is Bad Science. Computational Linguistics 33 (1): pp. 147–151.
17. *Ljubešić N., Erjavec T.* (2011), hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. Text, Speech and Dialogue 2011. Lecture Notes in Computer Science, Springer.
18. *Ljubešić N., Klubička F.* (2014): {bs,hr,sr} WaC—Web corpora of Bosnian, Croatian and Serbian. Proceedings of the 9th Web as Corpus Workshop (WaC-9). Gothenburg, Sweden.
19. *Markov A. A.* (1913), An Example of statistical research on the text of Eugene Onegin illustrated trial relations in a chain [Primer statisticheskogo issledovaniya nad tekstom "Yevgeniya Onegina", illustrirujuscikh svyaz' ispytaniy v tsepi], Imperial St. Petersburg Academy of Sciences Transactions [Izvestiya Inperatorskoy Akademii Nauk S.-Peterburga], series VI, vol. VII, pp. 153–162.
20. *Michelfeit J., Pomikálek J., Suchomel V.* (2014), Text Tokenisation Using unitok. In Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2014, pp. 71–75, 2014. Brno: NLP Consulting 2014.
21. *Moon, R.* (1998), Fixd Expressions and Idioms in English. A Corpus-Based Approach. Oxford: Clarendon Press.
22. *Piotrovskiy R. G.* (1968), Information measuring in language [Informatsionnye izmereniya yazyka], Leningrad.
23. *Pomikálek J.* (2011), Removing Boilerplate and Duplicate Content from Web Corpora. Ph.D. thesis, Masaryk University, Brno.
24. *Schäfer R., Bildhauer F.* (2012), Building Large Corpora from the Web Using a New Efficient Tool Chain. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12).
25. *Schäfer R.* (2015), Processing and querying large web corpora with the COW14 architecture. In: Proceedings of Challenges in the Management of Large Corpora (CMLC-3). Talk at Challenges in the Management of Large Corpora (CMLC-3) on July 20, 2015 in Lancaster.
26. *Schmid, H.* (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing. Manchester.
27. *Sharoff S.* (2006), Creating General-Purpose Corpora Using Automated Search Engine Queries. In: WaCky! Working Papers on the Web as Corpus. ISBN 88-6027-004-9, Bologna: Gedit Edizioni, pp. 63–98.
28. *Suchomel V., Pomikálek J.* (2012), Efficient Web Crawling for Large Text Corpora. In: Adam Kilgarriff, Serge Sharoff. Proceedings of the seventh Web as Corpus Workshop (WAC7). Lyon, 2012. pp. 39–43.

29. *Zakharov V. P., Masevich A. Ts.* (2014), Diachronic researches on the base of the Russian Google books Ngram Viewer text corpus [Diakhronicheskiye issledovaniya na osnove korpusa russkikh tekstov Google books Ngram Viewer], Srtuctural and Applied Linguistics [Strukturnaya i prikladnaya lingvistika], vol. 10, Saint-Petersburg, pp. 303–327.

THE BEGINNING OF A BEAUTIFUL FRIENDSHIP: RULE-BASED AND STATISTICAL ANALYSIS OF MIDDLE RUSSIAN

Berdičevskis A. (aleksandrs.berdicevskis@uit.no),
Eckhoff H. (hanne.m.eckhoff@uit.no)
UiT The Arctic University of Norway, Tromsø, Norway

Gavrilova T. (tanya96gavrilova@gmail.com)
The National Research University "Higher School of Economics", Moscow, Russia

We describe and compare two tools for processing Middle Russian texts. Both tools provide lemmatization, part-of-speech and morphological annotation. One ("RNC") was developed for annotating texts in the Russian National Corpus and is rule-based. The other one ("TOROT") is being used for annotating the eponymous corpus and is statistical. We apply the two analyzers to the same Middle Russian text and then compare their outputs with high-quality manual annotation. Since the analyzers use different annotation schemes and spelling principles, we have to harmonize their outputs before we can compare them. The comparison shows that TOROT performs considerably better than RNC (lemmatization 69.8% vs. 47.3%, part of speech 89.5% vs. 54.2%, morphology 81.5% vs. 16.7%). If, however, we limit the evaluation set only to those tokens for which the analyzers provide a guess and in addition consider the RNC response correct if one of the multiple guesses it provides is correct, the numbers become comparable (88.5% vs. 91.9%, 93.9% vs. 95.2%, 81.5% vs. 86.8%). We develop a simple procedure which boosts TOROT lemmatization accuracy by 8.7% by using RNC lemma guesses when TOROT fails to provide one and matching them against the existing TOROT lemma database. We conclude that a statistical analyzer (trained on a large material) can deal with non-standardised historical texts better than a rule-based one. Still, it is possible to make the analyzers collaborate, boosting the performance of the superior one.

Key words: Old Russian; Middle Russian; morphological tagging; lemmatization; rule-based approach; statistical approach

НАЧАЛО ПРЕКРАСНОЙ ДРУЖБЫ: ПРАВИЛОВЫЙ И СТАТИСТИЧЕСКИЙ АНАЛИЗ СТАРОРУССКОГО ЯЗЫКА

Бердичевский А. (aleksandrs.berdicevskis@uit.no),
Экхофф Х. (hanne.m.eckhoff@uit.no)
Университет Тромсё — Норвежский арктический университет, Тромсё, Норвегия

Гаврилова Т. (tanya96gavrilova@gmail.com)
Национальный исследовательский университет «Высшая школа экономики», Москва, Россия

Ключевые слова: древнерусский; старорусский; автоматическая морфологическая разметка; лемматизация; правиловый подход; статистический подход

0. Introduction

Apart from the usual challenges for NLP, processing of historical texts faces a number of additional ones, such as absence of a standard variant, absence of a standardized orthography and smaller resources, both in terms of existing tools and available texts (Piotrowski 2012). In this paper, we describe and compare two tools for processing Old/Middle Russian[1] texts. Both tools provide lemmatization, part of speech (POS) and morphological annotation. One analyzer (labeled "RNC"), described in Section 1, was developed for annotating parts of the historical subcorpus of the Russian National Corpus, and is rule-based. The other one (labeled "TOROT"), described in Section 2, is statistical, and is used for pre-annotating the eponymous corpus.

Since the analyzers were developed independently, and since they employ two different approaches, it is particularly interesting to compare their performance. Our expectation is that TOROT will perform better, since RNC does not perform disambiguation when several guesses are possible. Apart from testing this expectation empirically, we are also interested in checking whether it is possible to boost the TOROT performance by making the analyzers collaborate.

[1] For the purposes of this article we do not distinguish between Middle Russian "proper" and Church Slavic of the Russian recension, since both models deal well with both types of text, and since many texts are mixed. The text chosen for our performance test (see section 3.1) is a (late) Church Slavic text of the Russian recension.

1. The RNC analyzer

The "RNC analyzer" is a morphological analyzer for Middle Russian designed at Higher School of Economics (Moscow) for annotating the Middle Russian corpus (a part of the Russian National Corpus, ruscorpora.ru). The analyzer is based on Uniparser (Arxangelsky 2012, Arkhangelskiy, Belyaev and Vydrin 2012), which can give grammatical annotation to a text in any language provided that there exists a description of the language's grammar (a dictionary of inflections) and a grammatical dictionary of lexemes. The Uni-parser splits a word in all possible ways and looks for its parts in the description of the grammar and the grammatical dictionary. If one part of the word can be found in the dictionary of inflections, the other one in the dictionary of lexemes, and these parts are marked with the same inflectional class, then the word gets an analysis. There can be several possible analyses for one word. The parser does not create hypotheses for words which cannot be found in the dictionary and does not resolve ambiguity. The Uni-parser is intended for working on modern languages, so a module for dealing with spelling variability was developed by the third author of this paper. All letters which correspond to the same sound are reduced to one letter, geminate consonants are reduced to one letter, all jers between consonants are deleted and so on. Overall more than fifteen rules apply to a wordform before it is processed via Uni-parser.

The description of Middle Russian grammar was created manually. Due to the lack of a grammatical dictionary of Middle Russian, a grammatical dictionary of Old Church Slavic[2] (Poljakov 2014) was used. The dictionary was automatically adapted to Middle Russian: new inflectional classes were added, some regular differences between Old Church Slavic and Old Russian were taken into account. As far as Middle Russian contains both archaic and innovative forms, diachronic rules were applied to the dictionary. As a result, words which changed their inflectional class historically got two classes in the dictionary: the old one and the new one. Some word classes which are missing in the Old Church Slavic dictionary were added manually, e.g. pronouns and pronominal adjectives. The Uni-parser format requires information about all possible stems, so they were created automatically for each lexeme depending on its inflectional class. Different spelling variants were also added in the dictionary. For example, the lexeme "*княгиня*" 'princess' has two stems—"*княгин*" and "*княин*". The second one is a possible spelling variant with loss of the intervocalic *г*.

A lexical entry can contain several paradigms and several stems for each of them. For example the lexeme "*премощи*" 'overcome' has four stems in the dictionary (*премо*, *премог*, *преможе*, *премоз*). There can be up to fifteen stems for one lexeme.

[2] http://feb-web.ru/febupd/slavonic/dicgram/

2. The TOROT analyzer

The Tromsø Old Russian and OCS Treebank (TOROT, nestor.uit.no, see Eckhoff & Berdičevskis 2015) contains approximately 175,000 word tokens of annotated Old Russian and Middle Russian text (15th–17th century), fairly equally distributed between the two periods. The texts are all lemmatised and have fine-grained part-of-speech and morphology tags, in addition to syntactic annotation, yielding a large database of form, lemma and tag correspondences. This database is used systematically for linguistic preprocessing of texts: lemmatisation, part-of-speech assignment and morphological tagging. With a training set of this size, it is possible to train very successful statistical morphological taggers for these language stages, either separately or taken as a single stage. For this purpose, the TnT tagger (Trigrams 'n Tags, as described in Brants 2000), a statistical morphological tagger which takes trigrams and word-final letter sequences as its input is used (for the motivation behind this choice, see Skjærholt 2011).

To improve the performance of the tagger, both the training data and the new text to be tagged in the process are normalized. The normalisation consists in considerable orthographical simplification. All diacritics are stripped off, all capital letters are replaced with lower-case letters, all ligatures are resolved (e.g., ѿ to *от*), all variant representation of single sounds are reduced to one (all *о* variants are reduced to *о* and all *і* variants are reduced to *и*, for instance).[3] The juses are simplified to *я* and *у* (*ю*), and the jat to *е*.

When preprocessing a text, the tagger output is used in combination with direct lookups in the database.[4] For each word token in the text, it is checked whether that form is present in the database already, first as it is, then again with different kinds of orthographic simplifications. If one or more matches are found, the most frequent analysis (lemma + part of speech + morphology) is assigned. If the form is not found in the base, the TnT part-of-speech and morphology tag are assigned, and an attempt is made to find a suitable lemma in the database. If the word form (normalized to the lemma orthography style) matches a lemma with the part-of-speech tag the TnT tagger assigned, that lemma is assigned. If not, letters from the end of the word form are dropped one by one, the remainder checked again against the opening strings of lemmata of the correct part of speech. If no matches are found, a dummy lemma ("FIXME") is assigned, and the annotators will have to assign a lemma manually. This process is represented as a simplified flowchart on Figure 1.

---

[3] Supplementary materials can be found in the TROLLing data repository at http://hdl.handle.net/10037.1/10303. They include the normalization routine, the harmonization and comparison scripts (Section 3), and more detailed comparison results (Section 4).

[4] We are indebted to Professor Dag Haug at the University of Oslo for writing procedures for Latin and Greek, which we have modified for Slavic.

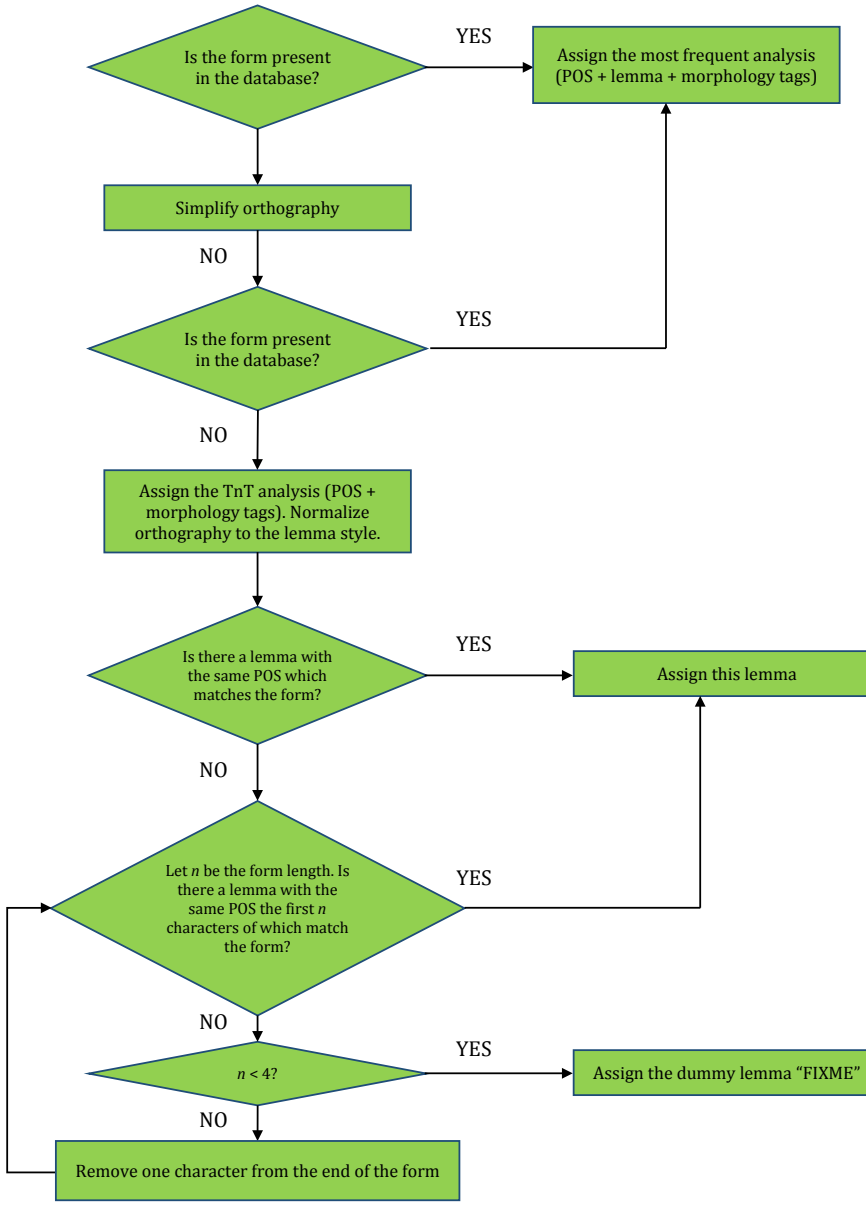


Figure 1. The TOROT automatic pre-annotation technique

3. Comparison

3.1. Test set and preprocessing

As a test set, we chose the preface to the "Life of Sergij of Radonezh" (1696 words in the unprocessed text), an early 15th century Russian Church Slavic text digitized after a late 16th century manuscript.[5] The text is late enough for the RNC analyzer (which is unlikely to perform optimally with earlier texts), but is still within the period which is of interest for TOROT. We normalized orthography (see Section 2) and ran both the RNC and TOROT analyzers on the otherwise unprocessed text.[6]

Since there were no discrete releases of the TOROT corpus at the time of the experiment (it is being expanded and corrected continuously), we preserved the training data as they were before the "Life of Sergij" was added to the corpus. That includes the whole set of Old and Middle Russian data that the TnT tagger was trained on (166,183 word tokens), and the full lemmata list that the TOROT analyzer used for lemma guessing (10,603 lemmata).

3.2. Gold standard and alignment

After preprocessing the annotation was manually corrected by a human expert, the annotation of every sentence was subsequently reviewed by at least one another expert. The resulting annotation was used as the gold standard.

The TOROT text import module assigns an id to every word in a text, and these ids are not normally changed by the annotators. These ids are used to align gold with the output of the TOROT analyzer. TOROT and RNC, in turn, are aligned using the source document: for every word in it, the corresponding TOROT guess and RNC guess (or a set of several guesses) are found. Note that TOROT always provides a guess (it may use the dummy lemma "FIXME", but the POS and morphology will still be provided), while RNC does not, which means that sometimes the TOROT guess will correspond to a blank.

Importantly, annotators can sometimes change tokenization, splitting an existing word token into two (14 cases, e.g. *неписано* > *не* 'not' and *писано* 'written'). This creates extra tokens (which are present in gold, but not in RNC or TOROT), so the total token count goes up to 1,710. Alternatively, the annotators can merge two tokens into one (5 cases, e.g. *во истинну* > *воистину* 'indeed'). This results in some tokens (*истинну*) existing only in RNC and TOROT, but not in gold. In both cases, there

---

[5] The text was digitised by Catherine Sykes and Hanne Eckhoff after the illuminated late 16th century manuscript of the Trinity Lavra of St. Sergius, available in facsimile online at http://old.stsl.ru/manuscripts/book.php?col=3&manuscript=001. The TOROT version is available at https://nestor.uit.no/sources/215.

[6] Note that the TOROT analyzer takes non-normalized text as input and uses both normalized and non-normalized tokens in the lookup process. We did an RNC analyzer test run with non-normalized input, the results were nearly the same.

always is at least one token matched against a blank, something which will be counted as an error for both analyzers.

3.3. Harmonization of the analyzers

We were interested in comparing the accuracy of POS tagging, full morphological tagging and lemmatization of the RNC and TOROT analyzers. Gold and TOROT, obviously, share the annotation format, but RNC uses a different one (different POS and morphological tags, lemmatization principles and orthography). In order to make the analyzers comparable, gold/TOROT and RNC have to be harmonized first. Some information is lost in the harmonization process, especially for the morphological tags.

3.3.1. Harmonization of the POS tags

The correspondences between RNC and TOROT POS classes are complicated. For every RNC tag, Table 1 lists all TOROT tags that can potentially correspond to it and were considered a correct match.

Table 1. POS tag correspondences

RNC POS tag	TOROT POS tag(s)
A	A-
A-PRO	A-, Pd, Pi, Pk, Pp, Pr, Ps, Pt, Px
ADV	Df
ADV-PRO	Df, Du
CONJ	C-, G-, Df
CONJ/PART	Df, G-
INTJ	I-
N	Nb, Ne
N-PRO	Pp, Pk, Pi, Px
NUM	Ma
PART	Df
PREP	R-
V	V-

If a token with a tag from the "RNC" column had one of the corresponding tags from the "TOROT" column in gold, the annotation was considered correct. TOROT tags: A- — adjective, Pd — demonstrative pronoun, Pi — interrogative pronoun, Pk — personal reflexive pronoun, Pp — personal pronoun, Pr — relative pronoun, Ps — possessive pronouns, Pt — possessive reflexive pronoun, Px — indefinite pronoun, Df — adverb, Du — interrogative adverb, C- — (coordinating) conjunction, G- — subjunction, Ma — cardinal numeral, I- — interjection, Nb — common noun, Ne — proper noun, R- — preposition, V- — verb. RNC tags: A — adjective, A-PRO — adjective pronoun, ADV — adverb, ADV-PRO — pronominal/interrogative adverb, CONJ — conjunction, CONJ/PART — a special tag for *да* 'so as / let", INTJ — interjection, N — noun, NUM — cardinal numeral, PART — particle, PREP — preposition, V — verb

3.3.2. Harmonization of the lemmatization

We consider lemmatization of a token correct if and only if both the lemma itself and the POS tag match the gold standard. There are numerous discrepancies in the spelling of the lemmata. TOROT consistently uses conservative orthography, largely following Sreznevskij (1895–1902) for the sake of better comparability of earlier and later texts. RNC focuses on the Middle Russian period and uses less archaic orthography. After manually analyzing the discrepancies, the following harmonization procedure was implemented. In gold lemmas (which are spelled according to the TOROT principles) all jers that are strong according to Havlik's law and the СъRC-rule were vocalized. Jers in the clusters *чьск* and *чьст* were vocalized, too. All remaining jers were deleted; yat was replaced by *е*; *кы/гы/хы* were changed to *ки/ги/хи*; double consonants were shortened to one. In RNC lemmas all jers were deleted; *зс* was changed to *сс*; double consonants were shortened to one; *о* was removed from *во-* and *со-* in the beginning of the word longer than four letters (this *о* is almost always a reflex of a jer in a prefix which gets missed by the vocalization rule applied to the gold lemmata); *жде* was changed to *же*. Ad hoc rules were created for three frequent lemmata: pronouns *сеи* and *тои* (changed into resp. *сии* and *тыи*) and verb *писати* (changed into *пьсати*).

After this procedure, the number of cases when a RNC lemmatization guess is unjustly labeled as wrong (some cases of "unexpected" jer vocalization; inconsistencies to the tagging of participles; some other spelling discrepancies) is reduced to 10, which we deem acceptable.

3.3.3. Harmonization of the morphology tags

The two morphological tag sets are not entirely compatible either. The RNC analyzer tags for a number of features that the TOROT analyzer ignores, namely transitivity (intr, tr), aspect (pf, ipf), reflexivity (med) and animacy (inan, anim).[7] In the comparison, these features are dropped. Both analyzers tag for long form/short form, but this is relevant for adjectives and participles only, and not adjectival pronouns. There are, however, considerable differences between the formats as to what is considered a pronoun and what an adjective. We therefore disregard this feature in the comparison. For the same reason, we ignore degree of comparison for adjectives and adverbs. Table 2 shows the harmonized tags per TOROT part of speech.

[7] We have nonetheless used the animacy tags to control for genitive-accusatives: TOROT tags these as genitives, RNC as accusatives. RNC masculine singular animate accusatives are thus considered matches of gold masculine singular genitives.

Table 2. Harmonized morphological tags used for comparison between the TOROT and RNC analyzers. The original RNC tags are in this format already, but are stripped of the features we chose to exclude (see main text). TOROT tags are converted into the simplified RNC format[8]

TOROT POS tag	Subcategory	Tagged for	Example of a possible harmonized tag
V-	l-participle	tense	perf
V-	participle	mood	participle[8]
V-	indicative	mood, tense, number, person	indic, praes, sg, 3p
V-	no mood feature	inflection	noninfl
V-	other	mood	inf
Nb, Ne	none	gender, number, case	f, sg, acc
A-, Pd, Pr, Ps, Pt	none	number, gender, case	sg, f, acc
Px, Ma, Mo	none	number, case	sg, acc
Pk, Pp, Pi	none	case	acc
Df	none	inflection	noninfl
Other	non-inflecting	inflection	noninfl

4. Results and performance boost

4.1. Results

The accuracy of lemmatization and POS tagging for TOROT and RNC are provided in resp. Tables 3 and 4. For RNC, we measure both "exact" (there is only one guess, and it is correct) and "fuzzy" (there are several guesses, and one of them correct) accuracy. Consider, for example, the form *paðu*. The RNC analyzer at its current stage will always assign three analyses to this form: the preposition *paðu* 'for the purpose of'; the verb *paðumu* 'take care' (2/3 person aorist singular); the adjective *paðъ* 'glad' (strong plural masculine nominative). Obviously, the RNC guess for *paðu* will never be an exact match. If, however, at least one of the three analyses correct, it will be considered a fuzzy match.

[8] In the vast majority of cases, the RNC analyzer is unable to provide a guess for participles, since the necessary rules have not been implemented yet. If it does hazard a guess, it is mostly erroneous. This tag is therefore simplified.

Table 3. Accuracy of the lemmatization and POS tagging by the TOROT analyzer

Metric	Lemma +POS, %	POS only, %	Number of tokens
Accuracy	69.8	89.5	1,710
Accuracy (when lemma is not "FIXME")	88.5	93.9	1,348
Accuracy (when RNC does not have a guess)	42.5	78.9	327

TOROT performs better on both accounts. Unsurprisingly, the numbers go up considerably for both analyzers if we take into account only those tokens for which they had a guess. RNC has a guess for 1383 tokens out of 1710 (81%). TOROT has a lemma guess for 1348 tokens (79%), a POS guess is always provided.

For RNC, fuzzy accuracy is much higher than exact one. When we are dealing only with tokens which have a guess, fuzzy accuracy is even higher than that of TOROT. Interestingly, if we limit ourselves to the tokens for which RNC failed to provide a guess, TOROT accuracy decreases noticeably. In other words, what is unsurmountable for RNC, is difficult for TOROT, too.

Table 4. Accuracy of the lemmatization and POS tagging by the RNC analyzer. "Exact" means that the analyzer provided a correct guess and nothing else; "fuzzy" means that there were several guesses, only one of each was correct

Metric	Lemma +POS, %	POS only, %	Number of tokens
Accuracy (exact)	47.3	54.2	1,710
Accuracy (fuzzy)	74.3	77.0	1,710
Accuracy (exact, when there is a guess)	58.5	67.0	1,383
Accuracy (fuzzy, when there is a guess)	91.9	95.2	1,383

A comparison of the morphological annotations is found in Table 5.

Table 5. Performance of the TOROT and RNC analyzers on morphological tags

	Accuracy, %	Number of tokens
TOROT	81.5	1,710
RNC (exact)	16.6	1,710
RNC (fuzzy)	70.2	1,710
RNC (exact, when there is a guess)	20.5	1,383
RNC (fuzzy, when there is a guess)	86.8	1,383

It should also be noted that a good number of the TOROT guesses are off by only one or two tags, as seen in Table 6. Since the TOROT morphological tags are 10-place positional tags, this can be measured by Hamming distances (the distance shows how many features got an incorrect tag).

The off-by-one errors are typically ambiguous forms such as *домъ* "house", which could be either nominative singular or accusative singular. It could also be a genitive plural, which might lead to a off-by-two error. Such morphological guesses are still of great practical use to the TOROT annotators, who will only have to make one or two corrections in the morphological tag, rather than providing a full new analysis.

Table 6. Hamming distances between gold tags and TOROT guess tags (10-place positional tag)

Hamming distance	count	%
0	16393	81.5
1	128	7.5
2	57	3.3
3	14	0.8
4	38	2.2
5	14	0.8
6	26	1.5
7	10	0.6
8	8	0.5
9	3	0.2
no tag	19	1.1

4.2. Boosting TOROT lemmatization accuracy

A question of practical importance is whether the analyzers are able to cooperate, helping each other out. Differences between the annotation formats, however, represent an important problem here. While we managed to harmonize the analyzers' outputs, some information got lost in the process. It does not seem realistic to do anything with morphological and POS tags, at least not without a more sophisticated harmonization. In addition, considering TOROT's better results, using it to boost RNC performance might be more complicated than simply using TOROT.

A promising avenue is to use RNC lemma guesses when TOROT fails to find one and resorts to "FIXME". We experiment with the following boosting procedure. For every token which is lemmatized as "FIXME" by TOROT and which has a RNC guess (either single or multiple), we go through all RNC lemma guesses. We harmonize the lemma and try to find a match in the (harmonized) TOROT lemma list (described in Section 3.1). If there is a match, the POS tag of the lemma guess and the potential match are compared, and if they are the same, the (non-harmonized version of the) lemma is taken as a guess,[9] otherwise the booster proceeds to the next RNC guess, if there is one. Obviously, this simple method can only work for lemmata which were

[9] Note that this can potentially result in a POS tag change due to the complex many-to-many corrrespondences used for the harmonization (see Table 1).

already in the TOROT list, but were not identified by the guesser described in Section 2. It transpires that even this can give performance a significant boost, see Table 7.

Table 7. Boosting TOROT lemmatization accuracy using RNC guesses

Metric	Lemma+ POS	POS only	Number of tokens
Success rate when fixing "FIXME"	90.3	92.7	165
Boosted TOROT accuracy	78.5	91.4	1,710

The booster attempts to provide lemma guesses for 165 tokens and gets it right in 149 cases. This increases TOROT lemmatization accuracy to 78.5% from 69.8% (see table 3). In addition, there is a slight improvement in POS tagging: 91.4% instead of 89.5%.

5. Conclusion

As was expected, the TOROT analyzer outperformed the RNC analyzer on all three accounts. There are several reasons for that.

The most prominent one is RNC's inability to disambiguate if there are several possible analyses. In addition, the selected text is non-standardized and displays considerable morphological variability and, even when consistent, idiosyncratic morphological endings, both in choice of form and orthography. The text also has numerous unresolved abbreviations. While our findings do not necessarily generalize to any historical text, these features are entirely typical of the texts of this era, and it seems reasonable to conclude that they strongly favour a statistical analyzer (trained on a large material) rather than a rule-based one. Furthermore, the RNC POS and morphological guesses are dependent on the analyzer's ability to come up with a lemma guess, whereas the TOROT analyzer guesses morphology with no reference to lemmatization. Finally, the RNC analyzer systematically misses a number of words altogether, such as all words with a titlo and most participles.

If we relax the evaluation criteria, requesting only the presence of a correct guess (not its uniqueness) and limit the evaluation set to those tokens for which RNC produces a guess, then RNC performs slightly better than TOROT. In other words, the analyzers are almost equally good at producing a guess, but differ in their ability to distinguish between several candidates. This finding shows that RNC has large potential, but one would have to develop a disambiguating technique in order to make this potential practically applicable, and this is a very time-consuming task.[10] At the current stage, the most practical thing to do if one wants to pre-annotate a Middle Russian text would be to use TOROT with the RNC lemmatization booster.

[10] Two anonymous reviewers asked how the RNC performance could be increased. Our answer is that the most important thing to do would be to implement disambiguation, but this task is far beyond the scope of this paper.

As described in 4.2, RNC can help TOROT out when it fails to provide a lemma guess. It is possible to check RNC lemma suggestions against the TOROT lemma list, and, if a match is discovered, use it as a lemma guess. This simple procedure boosts TOROT lemmatization accuracy by 8.7%, and POS tagging accuracy by 1.9%. For lemmatization, the difference is significant ($\chi^2(1) = 33.39$, $p < 0.001$), the effect size is small (Cohen's $h = 0.20$). For POS, the difference is not significant, the effect size is negligible ($\chi^2(1) = 3.46$, $p = 0.062$, $h = 0.07$). Thus, although a statistical model seems best for POS and morphological tagging, a rule-based model may considerably aid lemmatization.

Further work will no doubt result in better analyzers for Old and Middle Russian. However, the current approach is of great practical use. Especially for Middle Russian, there is a vast bulk of text available that could provide very interesting data for linguistic studies: the RNC Middle Russian subcorpus holds more than 7 million word tokens. Needless to say, the cost of manually analysis of all this text would be very high. On this background, an analyzer with around 80% success rate for both lemmatization and morphological annotation is a considerable gain, especially taking into consideration the unruly and unnormalized nature of these texts.

References

1. *Arxangelsky, T.* (2012), Principles of Morphological Parser Construction for Multi-structural Languages [Principy postroenija morfologičeskogo parsera dlja raznostrukturnyx jazykov]. PhD dissertation, Moscow State University.
2. *Arkhangelskiy T., Belyaev O., Vydrin A.* (2012), The creation of large-scaled annotated corpora of minority languages using UniParser and the EANC platform. Proceedings of COLING 2012: Posters. Mumbai, pp. 83–91.
3. *Brants, T.* (2000), TnT: a statistical part-of-speech tagger. In S. Nirenburg (ed.): Proceedings of the sixth conference on applied natural language processing 3, ANLC '00. Stroudsburg: Association for Computational Linguistics, pp. 224–231.
4. *Eckhoff, H. M., Berdičevskis, A.* (2015), Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank. Scripta & e-Scripta Vol. 14–15.
5. *Piotrowski, M.* (2012): Natural Language Processing for Historical Texts. Morgan & Claypool Publishers.
6. *Poljakov, A.* (2014), Church Slavonic corpus: spelling and grammar problems [Корпус церковнославянских текстов: проблемы орфографии и грамматики]. Przegląd wschodnioeuropejski Vol. 5 (1): 245–254.
7. *Skjærholt, A.* (2011), More, faster: Accelerated corpus annotation with statistical taggers. Journal for Language Technology and Computational Linguistics, Vol. 26:2.
8. *Sreznevskij, I.* (1895–1902), Materials for a Dictionary of the Old Russian Language [Materialy dlja slovarja drevnerusskogo jazyka]. Tipografija Imperatorskoj Akademii Nauk, St. Petersburg.

USING CONSTRAINTS ON A GENERAL KNOWLEDGE LEXICAL NETWORK FOR DOMAIN-SPECIFIC SEMANTIC RELATION EXTRACTION AND MODELING

Clairet N. (clairet@lirmm.fr)[1,2,3],
Ramadier L. (ramadier@lirmm.fr)[1,4],
Lafourcade M. (mathieu@lafourcade@lirmm.fr)[1]

[1]Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), Montpellier, France

[2]Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), Paris France

[3]Lingua et Machina, Boulogne-Billancourt, France

[4]IMAIOS, 672, Montpellier, France

We introduce a pattern-based approach applied to the semantic relation retrieval and semantic modeling. Our method relies upon the use of a general knowledge lexical semantic network built, shaped, and handled by crowdsourcing and GWAPs (games with a purpose). Implementing constraints on semantic relations available in the network increases the efficiency of the relation extraction process but also opens a semantic modeling perspective. In terms of (mostly horizontal) relation extraction, we tested our method on radiology reports in French. Our results show the interest of using a general knowledge lexical semantic network for the domain specific textual analysis as well as the interest of implementing series of constraints on semantic relations for the relation retrieval. We recently turned to the analysis of cooking recipes that stand for examples of domain specific instructional texts. Thus, in addition to the semantic relation discovery, we are building a method for the semantic modeling and conceptualization of cooking instructions. Its first results are presented below. Today, our results are available for French but we target extending the lexical network coverage to other languages in the next few years.

Keywords: semantic relation retrieval, semantic modeling, domain specific raw text analysis, lexical-semantic network

ИСПОЛЬЗОВАНИЕ ОГРАНИЧЕНИЙ В ЛЕКСИКО-СЕМАНТИЧЕСКОЙ СЕТИ ДЛЯ ИЗВЛЕЧЕНИЯ СЕМАНТИЧЕСКИХ ОТНОШЕНИЙ ИЗ СПЕЦИАЛИЗИРОВАННЫХ ТЕКСТОВ И СЕМАНТИЧЕСКОГО МОДЕЛИРОВАНИЯ

Клере Н. (clairet@lirmm.fr)[1,2,3],
Рамадье Л. (ramadier@lirmm.fr)[1,4],
Лафуркад М. (mathieu.lafourcade@lirmm.fr)[1]

[1]Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), Монпелье, Франция

[2]Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), Париж, Франция

[3]Lingua et Machina, Boulogne-Billancourt, Франция

[4]IMAIOS, Монпелье, Франция

Мы предлагаем подход к извлечению семантических отношений из неструктурированных текстов и семантическому моделированию основанный на использовании лексико-семантической сети общих знаний и семантических шаблонов. Используемая лексико-семантическая сеть развивается и поддерживается пользователями при помощи целевых онлайн игр и прямого участия (краудсорсинга). Специфика её структуры и применение ограничений (правил) к дугам сети повышают эффективность анализа семантики текста, а также открывают новые перспективы для (полу)автоматического семантического моделирования и концептуализации. Мы протестировали наш метод на материале корпуса рентгенологических заключений на французском языке с целью извлечения неиерархических отношений. Параллельно с исследованием этой проблематики, мы используем нашу методику для разработки модели кулинарного рецепта с целью концептуализации так как данный материал является примером процедурального текста. В настоящее время наши результаты касаются французского языка но мы планируем расширить нашу лексико-семантическую сеть включив в неё другие языки и, вместе с ней, возможности семантического анализатора.

Ключевые слова: семантический анализ, извлечение неиерархических семантических отношений, семантическое моделирование, анализ неструктурированного текста

Introduction

In recent years, the semantic analysis of the domain specific texts has been conducted on the basis of specific resources often without using any general knowledge repository. Instead of building a specific resource for our analysis, we immersed the domain specific knowledge into a general knowledge lexical semantic network. Then, we added constraints on the relations present in the lexical semantic network which improved our relation extraction results. Finally, we targeted an instructional text perspective and extend our approach to the modeling the sequences of actions corresponding to the cooking instructions. The paper is structured as follows. First, we give an overview of the state of the art and detail the resource we use for our experiments. Second, we introduce the IMAIOS system for semantic relation extraction and its results. Finally, we propose the MAKI system for the instructional text analysis and semantic modeling.

1. State of the art

In the literature, the "semantic relation" term corresponds to a variety of definitions. In the lexicographic perspective ex. Wordnet (Fellbaum, 1998), semantic relations are above all taxonomy relations (hyperonymy, hyponymy, meronymy). The semantic role approach understands semantic relations as roles handled by the terms in a context as proposed by (Dong et al., 2006) and described by (Morris and Hirst, 2004). Lexico-semantic networks such as BabelNet (Navigli and Ponzetto, 2012), ConceptNet (Liu and Singh., 2004), and JeuxDeMots (Lafourcade, 2007)[1] implement this kind of approach.

Most research work concerning the extraction of semantic relations focus on domain-independent relations (Snow et al., 2006; Chklovski and Pantel., 2004). In the paradigm of information retrieval, open information extraction systems such as (Banko and al., 2009) are also able to retrieve unknown relations. In the biomedical domain, there are four main techniques for relation extraction: finding co-occurrences (Jelier et al., 2005), using patterns or rules (Auger et al., 2008; Song et al., 2015; Rindflesch et al., 2000), implementing supervised learning-based approaches (Song et al., 2015; Rink et al., 2011), and using hybrid approaches (Suchanek et al., 2006; Chowdhury et al., 2012). The relation extraction between verbs (based on their arguments) yielded a number of methods. In particular, those of (Brody, 2007) and (Chambers and Jurafsky, 2008) focus respectively on arguments' role and discourse relations and the real order of actions instead of following the textual order.

[1] As a GWAP, JeuxDeMots includes a number of games for lexical acquisition and validation. The main word game involves two players who are asked questions to populate some knowledge type (ex. *What is the typical characteristic of cake*?) their answers are recorded and compared, the terms that appear in both answers are validated and either they are added into the network or their weight is augmented. More details are available in (Lafourcade, 2007).

The approaches to the semantic analysis of the cooking recipes fall into two major trends. The first one is centered on the concept of aliment, its possible features and adaptation perspectives. The IBM Chef Watson system implements this kind of approach. The second trend is related to the case-based reasoning paradigm. It uses cooking recipes as examples of instructional texts with the scope of analyzing and modeling work-flows. The Taaable (Badra et al., 2008) project is being developed in this perspective. Cooking instructions can be considered as short texts such as defined by (Pedersen, 2008[2]). Their analysis can also be conducted in the *mapping instructions to actions* perspective. In this paradigm the instructional text can be represented in formal language (Chen and Mooney, 2011), approahced from the machine learning linear policy estimation (Vogel and Jurafsky, 2010) in particular semantic role labeling (Malmaud et al., 2014) or alignment-based compositional semantics (Andreas and Klein, 2014) perspectives. In the knowledge engineering domain, a number of dedicated resources and models have been developed for food and nutrition: BBC Food Ontology[3], PIPS Food ontology[4], SOUR CREAM (Tasse and Smith, 2008) etc.

2. Crowd-sourced lexical semantic network

The lexical semantic network JeuxDeMots (Lafourcade, 2007) is an oriented, typed, weighted graph that contains 40M arcs (relations) linking 800K nodes. Unlike some of the similar resources (such as Wiktionnary), the JeuxdeMots graph features more than 100 relation types and includes various kinds of lexical, morphological, and semantic information. Therefore, it is relevant for the mining of horizontal relations such as *location*, *part-of*, *synonym*, causal and temporal relations, *characteristic*, *manner*, and more. The JeuxDeMots graph also implements inference and annotation schemes. The inference scheme (Zarrouk, 2015) for the graph semantic relations spreading introduces deductive, inductive, abductive and refinement approaches. The deduction and the induction mechanisms test the assumption of the transitivity of the is-a relation and use the logical blocking in case of polysemy. The blocking scheme is based upon the refinement concept and the annotation process applied to the premisses (typed and weighted outgoing relations inferred on the basis of hyperonyms/hyponyms of a term). Its final stage is the validation processed by a human expert. The refinement corresponds to the "real life" use of a term and may stand quite far from its lexicographic representation. This feature of our resource helps untangling some knotty problems such as the multi-word term detection and the polysemy management while analyzing raw text. The abduction scheme uses

---

[2] "A short written context consists of one to approximately 200 words of text that is presented to a human reader as a coherent source of information from which a conclusion can be drawn or an action taken"(Pedersen, 2008).

[3] http://www.bbc.co.uk/ontologies/fo/1.1

[4] http://cordis.europa.eu/project/rcn/71245en.html

synonym and similarity relations in order to infer new relations in the graph. To benefit from the inference scheme, a term is supposed to have at least one hyperonym, hyponym, synonym or refinement. The annotation scheme (Ramadier et al., 2014) allows defining and spreading the annotations over the relations already existing in the graph. It amplifies the inference scheme described above without adding new relation types into the graph.

The features of our resource grant the use of robust and somehow crude mining algorithms. The identification of compound terms can be made upstream by comparing the entities found in the text to the JeuxDeMots network. We use underscore to aggregate the two parts of a compound word. Thus, it is considered as an entity by the extractor (tibia_fracture). The polysemy resolution is based on the refinement. In the graph, all the refinements are linked to the core term. The disambiguation process is based on a triangular scheme including the term to disambiguate, its context and its refinements. The context is compared to the term refinements found in the graph. If a refinement do have some relation with the context (has at least one ingoing or outgoing relation to/from the context terms and these relation(s) has a weight $w > 50$) it is retained as it expresses the sense close to the context. The weight of a relation corresponds to the force of association between two words: how many crowdsourcers figured out the second term while considering the first one.

3. Extracting semantic relations from domain specific unstructured text: the IMAIOS system

The relation extraction approach implements series of semantic patterns. We understand semantic patterns as linguistic patterns similar to (Embarek and al, 2008) coupled with series of constraints on the relations of the JeuxDeMots graph. In the scope of the radiology report analysis and indexation and after being advised by radiologists, we have chosen 15 semantic relations relevant within this domain. These are in particular *r_isa* (generic terms), *r_synonym* (synonyms or quasi synonyms), *r_carac* (typical characteristics), *r_location* (typical location), *r_target* (disease target such as social group, organ), *r_part_of* (typical parts), *r_cause* (typical causes) etc. These relations can be of any general purpose. Some authors have already noticed that the use of patterns is an effective method for automatic information extraction from corpora if they are efficiently designed (Embarek et al., 2008; Cimino et al., 1993). For each relation type, we build patterns and match them with the sentences to identify the correct relation. These patterns are (for now) manually built through partial analysis of our corpus. In our experiment, we restricted ourselves to 42 semantic patterns, 12 of which are specific to medicine.

Table 1. Semantic pattern examples

Relations	Examples of patterns in English	Exemples of patterns in French
location	E1 on the level of E2	E1 au niveau de E2
location	E1 in E2	E1 dans E2
location	E1 is in the E2	E1 se trouve dans E2
location	E1 passing by E2	E1 passant par E2
causes	E1 may trigger E2	E1 déclenchant E2
characteristic	E1 is characterized by E2	E1 est caractérisé par E2
characteristic	Noun Adj	Nom + Adjectif
synonym	E1 also called E2	E1 encore appelé
causes	E1 can produce E2	E1 peut produire E2
consequence	E1 causes E2	E1 provoque E2
hyperonym	E1 is a E2	E1 est un E2
consequence	E1 leading to a E2	E1 menant à E2
target	E1 touching E2	E1 touchant E2
treatament	E1 treated by E2	E1 traité par E2
clinical sign	E1 accompanied by E2	E1 accompagné par E2

For some of the relations listed above, we encountered difficulties related to the ambiguity issue. For the location relation, we can distinguish two kinds of possible semantic relations depending on the pattern. The first pattern refers to the *r_location* relation (hepatocellular carcinoma is at the level of the liver). The second relation is holonomy (femur *r_holo* lower limb). For some connectors (of in caudate lobe of liver) both relations are correct (caudate lobe *r_location* liver and caudate lobe *r_holo* liver). We also make use of immediate co-occurences of entities for characteristic relation. For instance mutifocal hepatocellular carcinoma (HCC) appears five times together, so we consider multifocal as a probable characteristic of HCC (HCC *r_characteristic* multifocal).

Some linguistic patterns are inexpressive and it is hardly possible to determine the kind of the associated relation (ex. the french connector *de*, "of, from, because of"). Thus, we have added some semantic constraints on linguistic patterns. A semantic constraint is a condition that should verify the reification of one of variable of the pattern. There may be any number of constraints on $x and $y. Basically, a semantic constraint is a rule defined as follows: "if x is related to B than x is related to C" or x, Rb(x, B) =>Rc(x,C).

Table 2. Semantic pattern structure

Semantic pattern structure	Example
Lexical pattern (linguistic scheme)	"$x de $y"
Premisses (conditions or semantic rules)	*If $x r_isa illness & $y r_isa anatomical_location*
=> Conclusion (action)	$x r_location $y

To apply the semantic constraints' principle over our corpus and extract semantic relations, we use the following algorithm:

Let *S* the result set, being the empty set at initialization
Finding pattern occurrence in the text by moving a word window of size *n*
For each pattern occurrence applying constraints to the instantiated variables
If constrains are verified then the associated semantic relation is associated to $x and $y, that is to say added to *S*
Return S

From a corpus of more 30,000 medical reports, we extracted a random subset of around 120,000 relation instances for the different relation types[5]. About 800 of these relations were manually checked for evaluating precision. For assessing recall, we manually identified the relations in about 300 medical reports. Then we applied our algorithm.

Table 3. The IMAIOS system: relation extraction results

Relations	*Precision w/o constraints*	Precision with constraints	Recall	*F-measure w/o constraints*	F-measure with constraints	Contribution of the IMAIOS method (F-score)
cause	*74%*	90%	60%	*66%*	72%	+6.0
consequence	*70%*	89%	62%	*63.4%*	73%	+9.6
location	*48%*	83%	40%	*43.6%*	54%	+10.4
treatment	*70%*	88%	60%	*64.6%*	71.3%	+6.7
part-of	*32%*	75%	30%	*31%*	42.9%	+11.9
target	*45%*	80%	40%	*42.4%*	53.3%	+10.9
characteristic	*60%*	88%	58%	*60%*	70%	+10.0
lieu	*45%*	86%	40%	*41.7%*	54.6%	+12.0

The IMAIOS system has also been applied to other corpora. For a corpus of 45,000 cooking recipes, 245,000 semantic relations have been extracted with a precision of 95% (manually evaluated on a sample of 755 relations). Furthermore, we extracted 789,000 relations for randomly Wikipedia pages with a precision of 92% (manually evaluated on a sample of 1,250 relations). Hyperonym extraction on Wikipedia articles has a precision of about 94%.

[5] Even though we target some of the relation types according to our objectives, our system can extract any of the relation types present in the JeuxDeMots network on the basis of appropriate semantic patterns: taxonomic (*isa*, *hypo*, *has-part*), predicative (*agent*, *patient*), horizontal (*location*, *place*, *action place*, *instrument*, *manner*, *cause*, *consequence*, **qualia structure** (Pustejovsky, 1995) inspired relations (telic role, agentive role), and more.

4. Instructional text analysis, semantic relation extraction, and modeling

The MAKI System focuses on the analysis of cooking recipes taken as examples of instructional text, the extraction of temporal relations, and the modeling of the sequences of actions. Its ultimate goal is to build a conceptualizing work-flow to discover the canonical recipe on the basis of its variants found in the texts of recipes. It extends the use of semantic patterns by introducing the rules with not only two but more variables.

> $\forall X^n \forall x$, *patient*$(X^n, x) \wedge$ *is-a* $(x,$ *aliment*$) \exists y$, *pos* $(y,$ *Nom*$) \wedge$ *carac*$(y, z) \wedge$ *pos*$(z,$ *Ver:PP*$) =>$ *consequence*$(X^n, y) \wedge$ *successeur_temps*(X^n, y) *cf. "for each transforming action of the instructional text there is a state such as the consequence and the temporal successor of this action".*

The graph corresponding to the representation of the cooking instructions as sequences of actions is a bipartite graph with two types of nodes: states and actions. Our analysis strategy prompts that of (Bonfante et al., 2010) and also that of (Poria et al., 2014). We move from a syntactic dependency surface representation that can be obtained by using a parser such as Bonsaï PCFG-LA parser and MElt (Denis et Sagot, 2009), or by using the JeuxDeMots graph which contains such information. For each segment of the text we build an oriented, typed acyclic graph such as

$$|G| \models \forall x \exists (y, z.\ (edg\ (x, y, r') \wedge edg\ (x, z, r'')).$$

From the linguistic point of view, the MAKI system focuses on the following phenomena (among other): predominance of predicative structures; monotony of the argumental field (from one textual segment to another, we rediscover the same arguments such as *patient*, *instrument*, *place*, *quantifier* etc.); adjectival value of the past participle forms visible when observing the characteristic semantic feature (ex. *légumes blanchis,* "blanched vegetables"). These observations corroborate the assumption that any recipe action is spatially situated (utensil, table, kitchen), transforming (each action is followed by some new state of ingredients (patients)), and temporary finite.

In terms of computation, we introduce the dynamic creation of entities (nodes and directed arcs). The general analysis process starts from a surface representation built using the syntactic relation typed as *succ* (successor). Then, a context free grammar introduces constraints on the JeuxDeMots relations. As the rules (semantic constraints) keep on being applied, the JeuxDeMots graph is browsed and the semantic relations between these nodes keep on being discovered. The reification mechanism defined by (Zarrouk, 2015) is implemented. In this scheme, the graphs of the variables are dynamically built, the nodes typed as "rules" are created and linked to the reifications of a corresponding relation. We extend this scheme as we allow the creation of nodes to type the graphs of variables which facilitates their comparison. The variable graphs are linked to generic alignment entities using the relation type *r_head*. The lexical shape of these generic entities is that of the domain key terms (and not conceptual entities specifically created for the analysis) present in the JeuxDeMots

lexical semantic network. Thus, each generic alignment entity has a syntactic and semantic behavior and we can benefit from the power of our resource and its range of relation types. Main entities of this kind are: *action*, *état* (state), *préparation*, *ingrédient*. Moreover, we use such elements as grammatical features (parts of speech, gender, number etc.), word order, antecedence and reference markers in the text.

For the modeling of sequences of actions, we consider the instructional graphs built during our recipe analysis. We assume that actions pertaining to the same sequence share a certain number of resources (arguments) such as *patient* (ingredient), *instrument*, *place* (utensils, recipients). Therefore, graphs that do not have any of such resources in common and do not have any similar nodes (synonyms or quasi-synonyms otherwise linked by the relations typed as *part-of, substance, place of action, location*) are not part of a same sequence hyper-graph. We also noticed that the shortest path to the resources or groups of resources in the graph could be an indicator of the order of actions in a sequence.

To test our modeling system, we applied our scheme to 1,500 cooking instructions selected according to their predicative structure (at least one predicate and one patient), grammatical correctness, and length (4 to 12 tokens). For the evaluation purposes, we manually extracted semantic relations from our corpus (we obtained 3,878 relations) and built the generic alignment entities (1,720 entities). Then, we compared our semantic parser's results to this reference. The actions behind the textual instructions have been modeled exactly in 57% of cases (855 graphs have been built with all the expected nodes and all the appropriate semantic relations). 28% of instructions have been partly modeled (which corresponds to 435 instructions modeled). Finally, in 14% of cases our system failed in building alignment hyper-graphs which is due to the need of improving the coverage of the Jeux DeMots graph (new relation types). The semantic relations we extracted and the entities we built according to the rules show the following results:

Table 4. The MAKI system: relation extraction and semantic modeling first results

Relation type	Reference	Extraction	Recall	Precision	F-score	Modeling	
patient	1,500	1,700	1	88%	94%	action	966
characteristic	505	570	94%	86%	90%	state	70
manner	378	288	50%	98%	66%	event	15
successor_time	420	67	78%	1	87%	set_of_ingredients	210
has-part	168	145	73%	1	84%	mixture ::	63
quantifier	84	77	83%	1	91%		
place	462	444	69%	88%	77%		
place of action	336	329	61%	82%	70%		
instrument	25	21	68%	1	81%		
agent (not expected)	-	205	-	-	-		
agent-1 (not expected)	-	200	-	-	-		
Totals	3,878	4,152	-	-	-	Total	1,324
average	-	-	**75%**	**93%**	**82%**	**Modeling accuracy**	**77%**

Conclusion

While the machine learning techniques are dominant in the research field of computational linguistics within the text analysis, the graph based approach with crafted knowledge remains a very promising area for fine semantic analysis of raw text as well as for the terminological and relation retrieval. Indeed, a common knowledge graph gives access to the extra-linguistic information which is not part of the text as a percept and which rarely appears in texts. Thus, this type of resource is relevant for the domain specific texts' analysis.

References

1. *Abacha, A. B., & Zweigenbaum, P.* (2011). Automatic extraction of semantic relations between medical entities: a rule based approach. J. Biomedical Semantics, 2(S-5), S4.
2. *Auger, A., & Barrière, C.* (2008) Pattern-based approaches to semantic relation extraction: A state-of-the-art. Terminology, 14(1), pp. 1–19.
3. *Badra, F., Bendaud, R., Bentebibel, R., Champin, P. Cohan, J., Cordier, A.* (2008) TAAABLE : Text Mining, ontology Engineering, and Hierarchical Claasification for Textual Case-Based Cooking. Actes de 9th European Conference on Case-Based Reasoning, 5239, 219–228.
4. *Banko, M., Michael J., Cafarella, M. J., Stephen Soderland, S., Broadhead, M., and Etzioni, O.* (2007) Open Information Extraction from the Web, International Joint Conference on Artificial Intelligence (IJCAI).
5. *Bodenreider, O.* (2004). The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32(suppl 1), D267–D270.
6. *Chambers, N., & Jurafsky, D.* (2008). Unsupervised learning of narrative event chains. Proceedings of the Association of Computational Linguistics, 31(14), 789–797.
7. *Chklovski, T. and Pantel, P.* (2004). Verbocean : Mining the web for fine-grained semantic verb relations. In Dekang Lin and Dekai Wu, editors, Proceedings of EMNLP 2004, 33–40, Barcelona, Spain, July. Association for Computational Linguistics.
8. *Chowdhury, M. F. M., & Lavelli, A.* (2012, April). Combining tree structures, flat features and patterns for biomedical relation extraction. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (pp. 420–429). Association for Computational Linguistics.
9. *Cimino, J. J., & Barnett, G. O.* (1993). Automatic knowledge acquisition from MEDLINE. Methods of information in medicine, 32, 120–120.
10. *Das, D., Chen, D., Martins, A., Schneider, N., and Smith, N.* (2014). Frame-semantic parsing. Computational Linguistics.
11. *Denis, P. and Sagot, B.* (2012). Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. Language Resources and Evaluation, (46): 721–736.
12. *Dong, Z., Dong Q.* (2006). HowNet and the Computation of Meaning. World Scientific Publishing.

13. *Dufour-Lussier, V., Le Ber, F., Lieber, J., & Nauer, E.* (2014). Automatic case acquisition from texts for process-oriented case-based reasoning, Information systems, 40, 153–167.
14. *Embarek, M., & Ferret, O.* (2008). Learning Patterns for Building Resources about Semantic Relations in the Medical Domain. In LREC, may 2008.
15. *Esuli, A., Marcheggiani, D., & Sebastiani, F.* (2013). An enhanced CRFs-based system for information extraction from radiology reports. Journal of biomedical informatics, 46(3), pp. 425–435.
16. *Fabre, C., Bourigault, D.* (2006). Extraction de relations sémantiques entre noms et verbes au-delà des liens morphologiques. TALN 2006, Leuven, 10–13 avril 2006, pp. 121–129.
17. *Fellbaum, C.* (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.
18. *Lafourcade, M.* (2007, December). Making people play for Lexical Acquisition with the JeuxDeMots prototype. In SNLP'07: 7th international symposium on natural language processing (p. 7).
19. *Lafourcade M.*(2011). Lexique et analyse sémantique de textes—structures, acquisitions, calculs, et jeux de mots, HDR, Université Montpellier II.
20. *Liu, H., Singh, P.* (2004). ConceptNet—a practical commonsense reasoning toolkit. BT technology journal, 22(4):211–226.
21. *Malmaud, J., Wagner, E. J. Chang, N., Murphy, K.* (2014). Cooking with Semantics, Actes de ACL 2014, 33–38.
22. *Morris, J., & Hirst, G.* (2004). Non-classical lexical semantic relations. Actes de Workshop on Computational Lexical Semantics, pp.46–51.
23. *Navigli, R., Ponzetto, S. P.* (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, 193, Elsevier, pp. 217–250.
24. *Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, K., Marinov, S., and Marsi, E.* (2007). Maltparser : A language-independent system for data-driven dependency parsing. Natural Language Engineering 13(2) :95–135.
25. *Pedersen, T.*(2008). Computational Approaches to Measuring the Similarity of Short Contexts : A Review of Applications and Methods. CoRR abs/0806.3787
26. *Poria, S., Agarwal, B., Gelbukh, A., Hussain, A., Howard, N.* (2014). Dependency-Based Semantic Parsing for Concept-Level Text Analysis, in Computational Linguistics and Intelligent Text Processing: 15th International Conference, CICLing 2014, Nepal, April 6–12, Springer Berlin Heidelberg :113–127.
27. *Pustejovsky, J.*(1995) The Generative Lexicon, MIT Press, Cambridge.
28. *Ramadier, L., & Zarrouk, M.* (2014). Annotations et inférences de relations dans un réseau lexico-sémantique :application à la radiologie, Actes de TALN, 103–112.
29. *Rindflesch, T. C., Bean, C. A., & Sneiderman, C. A.* (2000). Argument identification for arterial branching predications asserted in cardiac catheterization reports. In Proceedings of the AMIA Symposium (p. 704). American Medical Informatics Association.
30. *Rink, B., Harabagiu, S., & Roberts, K.* (2011). Automatic extraction of relations between medical concepts in clinical texts. Journal of the American Medical Informatics Association, 18(5), 594–600.

31. *Song, M., Kim, W. C., Lee, D., Heo, G. E., & Kang, K. Y.* (2015). PKDE4J: Entity and relation extraction for public knowledge discovery. Journal of biomedical informatics, 57, 320–332.
32. *Song, M., Yu, H., & Han, W. S.* (2011). Combining active learning and semi-supervised learning techniques to extract protein interaction sentences. BMC bioinformatics, 12(Suppl 12), S4.
33. *Snow, R., Jurafsky, D., & Ng, A. Y.* (2006, July). Semantic taxonomy induction from heterogenous evidence. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (pp. 801–808). Association for Computational Linguistics.
34. *Suchanek, F. M., Ifrim, G., & Weikum, G.* (2006, August). Combining linguistic and statistical analysis to extract relations from web documents. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 712–717). ACM.
35. *Tasse, D. and Smith, N.*(2008). SOUR CREAM: Toward Semantic Processing of Recipes. TechnicalReport CMU-LTI-08–005, Carnegie Mellon University, Pittsburgh, PA.
36. *Zarrouk, M.* (2015) Consolidation endogène de réseaux lexico-sémantiques : inférence et annotation des relations, règles d'inférence et langage dédié. Thèse de doctorat, Université Montpellier II.

ОЦЕНКА СТЕПЕНИ СВЯЗИ В СИНТАКСИЧЕСКИХ КОНСТРУКЦИЯХ С ИСПОЛЬЗОВАНИЕМ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ СЛОВ

Букия Г. Т. (gregorybookia@yandex.ru),
Протопопова Е. В. (protoev@yandex.ru),
Паничева П. В. (p.panicheva@spbu.ru),
Митрофанова О. А. (o.mitrofanova@spbu.ru)

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

В статье описан оригинальный подход к оценке связей в синтаксических конструкциях (прежде всего, в сочетаниях «прилагательное + существительное»). В экспериментах используются векторные модели, основанные на Word2Vec и на авторской мере оценки связей, учитывающих сочетания, не наблюдаемые в корпусе текстов. Исследовательские данные позволяют делать выводы о композиционности сочетаний и о синтаксических связях между частями сочетаний. Оценка параметров используемых нами моделей осуществляется в рамках так называемой псевдо-дизамбигуации. В ходе тестов обе модели показали высокие результаты. Мы провели анализ ошибочных разборов сочетаний и выявили несколько типов ошибок, в числе которых метафорические конструкции, конструкции с частично десемантизированными элементами, переходные случаи.

Ключевые слова: дистрибутивная семантика, композиционные сочетания, сочетания «существительное+прилагательное», меры ассоциации, Word2Vec, псевдо-дизамбигуация, русскоязычные корпусы

ESTIMATING SYNTAGMATIC ASSOCIATION STRENGTH USING DISTRIBUTIONAL WORD REPRESENTATIONS[1]

Bukia G. T. (gregorybookia@yandex.ru),
Protopopova E. V. (protoev@yandex.ru),
Panicheva P. V. (p.panicheva@spbu.ru),
Mitrofanova O. A. (o.mitrofanova@spbu.ru)

St. Petersburg State University, St. Petersburg, Russia

[1] The reported study is supported by RFBR grant № 16-06-00529 “Development of a linguistic toolkit for semantic analysis of Russian text corpora by statistical techniques”.

In the paper we present distributed vector space models based on word embeddings and a specific association-oriented count-based distributional algorithm which have been applied to measuring association strength in Russian syntagmatic relations (namely, between nouns and adjectives). We discuss the compositional properties of the vectors representing nouns, adjectives and adjective-noun compositions and propose two methods of detecting the syntactic association possibility. The accuracy of the proposed measures is evaluated by means of a pseudo-disambiguation test procedure and all models show considerably high results. The errors are manually annotated, and the model errors are classified in terms of their linguistic nature and compositionality features.

Keywords: distributional semantics, compositional collocations, adjective-noun phrases, association measures, Word2Vec, pseudo-disambiguation, Russian corpora

1. Introduction

Semantic compatibility (term used, for example, in (Ghomeshi, Massam 1994)) or the ability of two words or constructions to collocate has been widely studied within different theoretical frameworks (Goldberg 1995), (Apresjan 1974). The meaning of complex linguistic units such as collocations is generally assumed to be non-compositional so it is not fully derived from the meaning of their parts. Consider Apresjan's example of Russian adjectives 'goryachij' and 'zharkij' (Apresjan 2010), both translated in English as 'hot'. They are treated as synonyms, although in fact they are not fully interchangeable in contexts as they have virtually non-overlapping sets of collocates. 'Goryachij' is used to refer to a local sensation (*'goryachij shokolad'*—hot chocolate) and 'zharkiy' expresses the idea of a more general sensation of the environment (*'zharkoe leto'*—hot summer).

Such restrictions can be treated within Construction Grammar (Goldberg 1995) theory claiming that lexical constructions reveal the unity of form and meaning. The form is maintained by fixed elements of constructions on the one hand, and on the other hand, by selectional (morphosyntactic, lexical-semantic, propositional, etc.) restrictions imposed on the slot fillers. Construction Grammar observes a wide range of variations,from free co-occurrence of lexical features to highly idiomatic units. In our study constructions are treated as multilevel structures combining lexical (lemmata, wordforms), grammatical and semantic features. Such an approach allows us to describe collocability of a target word in a given sense in terms of constructions.

The degree of association in lexical constructions is an important factor in such NLP applications as paraphrase generation for machine translation, language modelling, automated and semi-automatic dictionary acquisition, semantic role labelling, word sense disambiguation, etc. A number of collocation extraction methods rely on corpus evidence assuming that if a construction occurs in texts, its components can be combined. However, these methods are not applicable when a pair of words is not observed in texts. Moreover, we can imagine occasional word combinations that are not generally expected in natural context (*'дремучее равновесие'*—primeval balance).

In our study we compare two approaches to measuring association strength in word combinations revealing possibility of syntagmatic relations, the first one assuming compositionality of their meaning, the second one implying that such word combinations have a meaningful unit which is not derived from word meanings. The paper is organized as follows: first of all, an outline of the research in the field is presented. Then, we describe two approaches to measuring association strength using distributed vector representations. Finally, the performance of the models is evaluated within a pseudo-disambiguation benchmark, and in conclusion a brief error analysis is presented.

2. Related work

Recently distributional semantic modelling has been applied to studying meaning of complex linguistic units (constructions, clauses, sentences) with the help of vector space models and their modifications (Kolb 2008), (Pekar, Staab 2003), (Sahlgren 2006), (Schütze 1992), (Widdows, Cohen 2010), etc. SemEval 2014 competition[2] included evaluation of compositional distributional semantic models of full sentences for English. One of the recent examples is COMPOSES which uses compositional operations to model linguistic units in semantic space. "Content" words (e.g. nouns) are represented as vectors, while relational words (e.g., adjectives) correspond to functions mapping input items to compositional structures (Baroni et al. 2014).

A survey on mathematical operations applied to determine compositional meaning is presented in (Kartsaklis, Sadrzadeh 2013). The authors focus their attention on tensor-based models where relational words (verbs, adjectives) are regarded as tensors. The distributed vector representations (Mikolov et al. 2013a) are also studied with respect to their compositionality (Mikolov et al. 2013b).

Distributional models for Russian have been applied in a number of applications. Serelex semantic model is incorporated in an information retrieval system which gets a target word as an input and gives a list of its associates as an output (Panchenko 2013). It provides contextual correlates for a target word which are ranked according to an original similarity measure based on lexical-syntactic patterns.

The evaluation of various association measures and Russian distributional models has been discussed in RUSSE competition (Panchenko et al. 2015). However, semantic relatedness evaluation involves only paradigmatic relations between lexical units. Thus, to our knowledge, there has been no evaluation of vector space models applied to syntagmatic relations in Russian.

A recent study concerning association strength measurement in syntactic constructions and testing methodology is described in (Bukia et al. 2015). The experiments are conducted on syntactic constructions. The authors train association measure on adjective-noun collocations from a very small corpus of 350 thousand sentences. The algorithm yields high performance in predicting association possibility although it is based on a small amount of training data. Their approach is detailed below and compared with association strength evaluation results produced by vector space modelling.

---

[2] http://alt.qcri.org/semeval2014/task1/

Association strength measurement is closely related to identification of abnormal lexical compositions (Vecchi et al. 2011) and automatic lexical error detection (Kochmar, Briscoe 2013). The latter work presents a number of semantic anomaly measures in a vector space. We adopt one of the measures and apply it to a semantic space with reduced dimensionality produced by Word2Vec.

3. Distributed word representations and their application to association measurement

3.1. Distributed word representations in word2vec toolkit

Continuous word representations in vector space have been gaining extreme popularity since (Mikolov et al. 2013a). As reported in the paper, high quality word vectors are obtained by training recurrent neural network with two different architectures—continuous-bag-of-words (CBOW) and skip-gram. Both yield considerable results in similarity and association measurement when using the cosine similarity measure. The authors implement their approach in a widely used word2vec[3] toolkit.

3.2. Distributed word vectors and linguistic regularities

(Mikolov 2013a) have proposed a questionnaire method (later elaborated in (Mikolov et al. 2013c)) to estimate word vector representations in terms of semantic and syntactic relationships between words that are learned automatically. The question is formed of two pairs of words with the same relationship such as "What is the word (*x*) that is similar to *small*(x_c) in the same sense as *biggest*(x_b) is similar to *big*(x_a)?" It turns out that the vector

$$y = x_b - x_a + x_c$$

is most similar to *x* in terms of cosine similarity. This was proved on several groups of questions including semantic relationship (*the capital of*, *the currency of*, *the female of*, etc.) and syntactic or, more precise, grammatical ones (*past form of*, *superlative form of*, etc.).

The difference of two word vectors characterizes their relation which is independent of their own meanings and can be used to infer the missing word in a different pair representing the same relation.

This observation may be applied to measuring association in syntactic constructions even for word-pairs not attested in the corpus. We assume that if a pair of words comprises a collocation or construction, there is a regular semantic relationship between the two words, which is systematically replicated in other word-pairs attested in the corpus. Thus we can find such a pair of words in the corpus that its difference vector is similar to the corresponding difference vector of the given words. Otherwise,

[3] https://code.google.com/archive/p/word2vec/

if the combination is unacceptable, the difference vector is unpredictable and does not have similar vectors in corpus. This difference vector often accounts for a relation which can not be formulated clearly but appears regularly in syntactic word combinations. Consider several examples of a noun + adjective combination and the nearest pair:

- 'овощной салат' (vegetable salad)—'конфетная коробка' (a box of sweets), 'гороховый суп' (pea soup);
- 'чёрный кофе' (black coffee)—'тёмное пиво' (dark beer), 'розовое мартини' (pink martini).

  The example sets consist of definitions by content and color respectively.

  The association measure for a combination (*a*,*n*) is formulated as:

$$W2V_{rel} = \max_{a_i,n_j \in K} \frac{< n, n_i - a_i + a >}{|n||n_i - a_i + a|},$$

where the maximum value is found over all pairs (a_i, n_j) occurring in the corpus. This measure is referred to as $W2V_{rel}$ below.

3.3. Compositional approach to association measurement

We adopt the compositional approach investigated in (Kochmar, Briscoe 2013), (Vecchi et al. 2011). The assumption is that the vector representing a noun-adjective composition is meaningful if it is closely related to the head of the composition, i.e. the initial noun. The similarity measure between the composition and the head noun is expected to positively reflect the acceptability of the noun-adjective association. The acceptable compositions are expected to be ranked as more similar to the initial head nouns than the unacceptable ones. However, with normalized vectors, as in Word2Vec approach, the monotonicity of the functions *Similarity1* and *Similarity2* is the same, although *Similarity2* measures the simple cosine similarity between noun and adjective:

$$Similarity1(noun, adj) = cos(noun + adj, adj)$$
$$Similarity2(noun, adj) = cos(noun, adj)$$

Quantifying similarity between the noun and the adjective in the same vector space yields here the same result as when quantifying similarity between the initial noun and the attributive noun phrase. The latter formula is linguistically motivated and naturally interpreted, which is not so obvious about the former. These values will be referred to as *Comp* below.

3.4. Count-based distributional approach

We compare the described methods to an approach proposed in (Bukia et al. 2015). Their association measure is based on distributional word properties concerning only a fixed construction, namely, the noun-adjective association (referred to as *D* below).

The basic assumption is that if two words relevant for a construction slot collocate in texts with similar words (contexts) relevant for another slot, the probability of the first target word to be combined with the contexts of the second target word and, vice versa, is high, even when some pairs are not observed in texts. This idea is formally expressed by the notion of confusion probability, which is computed as follows: given the contexts of the first word $c(x_1)$ and the second one $c(x_2)$, their confusion probability is equal to P:

$$\mathbb{P}\{x_1 \sim x_2\} = \frac{|c(x_1) \cap c(x_2)|^2}{|c(x_1)||c(x_2)|}$$

The association strength between two words in a collocation occurring in a corpus is usually computed by means of Fisher's exact test (Stefanowitsch 2003). The distributional association measure between a noun and an adjective in a collocation $D(a,n)$ is then defined as an average of such counts over all confusable words weighted by the confusion probability. As discussed in (Bukia et al. 2015), the highest results are produced by such a measure based on mutual information (MI) counts.

4. Evaluation

4.1. Data and experimental setup

We use a corpus of Russian fiction (146M tokens obtained from M. Moshkov's digital library, URL: lib.ru). All preprocessing (tokenization, lemmatization, shallow morphological analysis) was performed by means of PyMorphy2 Python library (URL: http://pymorphy2.readthedocs.org/en/latest/). About 157K (80K unique) adjective-noun pairs were extracted from these texts.

The 150-dimensional vectors were trained using Gensim library (Rehurek, Sojka 2010) with skip-gram architecture and 4-word window. The count-based distributional association takes into account only corpus frequencies of a noun, an adjective and their combination. The evaluation procedure follows pseudo-disambiguation test as described in (Bukia et al. 2015). It has been also used in (Pekar 2004), (Tian et al. 2013).

The following lemmata were extracted from the corpus:

- 500 random nouns $N = \{n_i\}$;
- for each noun a random adjective a_i collocating with this noun;
- for each a the nearest adjective by frequency $X = \{x_i\}$ (not attested in combination with the corresponding noun n_i).

All combinations (a_i,n_j) are then removed and the system is trained on the rest of the corpus. Thus, 500 triplets consisting of a target noun, an acceptable and an unacceptable combination are obtained. The task is to find out, which combination of an adjective and a noun was removed, i.e. which one is acceptable but deleted from the final training corpus. In our case, the first association value is expected to be higher than the second one.

4.2. Results

It should be noticed that the pseudo-disambiguation task is limited by the fact that we do not know anything about the second combination not attested in the corpus. Thus, the results were manually checked in order to eliminate malformed triplets. In some cases, either both combinations are acceptable or even none of the chosen ones.

The accuracy counts are presented in Table 1 in the following order:

- raw result based on the assumption that the first possible pair is acceptable while the second one is not (*Acc*);
- pseudo-disambiguation accuracy calculated after manual annotation of triplets (*Corr*).

As mentioned above, the models are denoted as follows:

- $W2V_{rel}$ for vector difference based measure (Section 3.2);
- *Comp* for measure based on vector composition (Section 3.3);
- *D* for a simple distributional measure (Section 3.4).

First of all, it should be noticed that the models based on word embeddings achieve higher accuracy than a count-based one. However, the latter one has an important advantage: its results are easy to implement and interpret.

Secondly, the results presented below should be compared only with the data provided by the models performing the same task: estimating association strength for unseen combinations. The most recent work (Tian et al. 2013) based on quite different principles reports 88% accuracy. Thus, the discussed models yield state-of-the art performance.

Table 1. Accuracy and real error percentage in the pseudo-disambiguation task

	$W2V_{rel}$	*Comp*	*D*
Acc	76%	81%	75%
Corr	88%	93%	84%

4.3. Error analysis

After manual error annotation the errors of different models are compared. Common errors, i.e. shared by all three methods, can be divided into two groups concerning their source: acceptable combinations representing rare or occasional metaphorical expressions (*'информационная чума'—information boom*, *'круглая сирота'—a total* (literally *'round'*) *orphan*, etc.) and those containing a word with a vague or general meaning (*'европейский квартал'—european quarter*, *'серьезное ухаживание'—earnest courting*, etc.).

The rest of the errors, i.e. the model-specific ones, were ordered by the acceptable combination frequency. In each experiment a group of very rare (acceptable) combinations can be found: *'суеверный закон'—superstitious law*, *'безработный фанатик'—unemployed fanatic*. These expressions are either rare themselves or contain a rare word, and even a native speaker is scarcely able to construct a sentence where

these combinations are justified. The top combinations are constructions in the sense of Construction Grammar (Goldberg 1995): their meaning is not additive and can not be simply derived from the meaning of the constituents: *'жевательная резинка'—chewing gum*, *'трезвая голова'—reasonable person* (literally *'sober head'*).

The middle of these ordered lists contains real errors which are due to an algorithm structure or its assumptions: *'копировальный центр'—copy center*, *'сумасшедшая история'—mad story*. These combinations are less idiomatic and are constructed regularly. The Word2Vec compositional similarity measure (section 3.3) fails to extract such combinations because the constituents appear to have too few intersecting contexts. The errors of the count-based distributional model may also be explained by the underlying assumption that similar words occur in similar noun-adjective contexts. On the other hand, such expressions are correctly processed by the Word2Vec relative measure (see section 3.2), meaning that analogous regular relations between attested nouns and adjectives were observed when looking for the nearest difference vector. Several examples are presented in table 2.

Table 2. Examples of nearest difference vector combinations

test combination	nearest difference combination
жевательная резинка—chewing gum	*кавказский хребет—Caucasian chain* *цементная ступенька—cement step*
копировальный центр—copy center	*патрульный корабль—patrol ship*

Finally, it should be mentioned that although Word2Vec compositional measure has shown the best results, it can not be improved, because its assumption is not applicable in all real world cases. The accuracy of Word2Vec relative measure, on the contrary, can be increased, since the core idea of an additional meaning can be observed in real data.

5. Conclusion

We have applied the task of measuring association strength between Russian nouns and adjectives to compare compositional and relative Word2Vec semantic models with a simple distributional association measure. The test was conducted following a conventional pseudo-disambiguation methodology. The models were trained with a 11M sentences corpus where all in-sentence co-occurrences of the word pairs are deleted.

Both measures based on Word2Vec models outperformed a simpler count-based one and achieved state-of-the-art accuracy. The error analysis allows us to talk about future improvements by applying a more sophisticated measure to determine a syntagmatic relation. More exactly, we are going to focus on the interpretation of the difference vector. Another important concern is the testing methodology which is also subject to future investigation and improvement based on human judgements. For example, separate datasets for compositional and idiomatic combinations should be created and manually assessed.

References

1. *Apresjan, Ju. D.* (1974), Leksicheskaja semantika. [Lexical semantics], Moscow, Nauka.
2. *Apresjan, Ju. D.* (ed.). (2010), Prospekt aktivnogo slovarya russkogo jazyka [The prospect of active Russian dictionary)], Moscow.
3. *Baroni, M., Bernardi, R., Zamparelli, R.* (2014), Frege in space: A program of compositional distributional semantics, Linguistic Issues in Language Technology, Vol. 9, CSLI Publications.
4. *Biemann, C.* (2007), Unsupervised and knowledge-free natural language processing in the structure discovery paradigm, PhD thesis, University of Leipzig.
5. *Bukia, G., Protopopova, E., Mitrofanova, O.* (2015), A corpus-driven estimation of association strength in lexical constructions, Sergey Balandin, T. T., Trifonova, U.(eds.), Proceedings of the AINL-ISMW FRUCT, FRUCT Oy, Finland, pp. 147–152, http://fruct.org/publications/ainl-abstract/files/Buk.pdf
6. *Ghomeshi, J., Massam D.* (1994), Lexical/syntactic relations without projection, Linguistic Analysis, Vol. 24, Issues 3–4, pp. 175–217.
7. *Goldberg, A.* (1995), Constructions: A construction grammar approach to argument structure, University of Chicago Press, Chicago.
8. *Kartsaklis, D., Sadrzadeh, M., et al.* (2013), Prior disambiguation of word tensors for constructing sentence vectors, Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNL), pages 1590–1601, Seattle, USA.
9. *Kochmar, E., Briscoe, T.* (2013), Capturing anomalies in the choice of content words in compositional distributional semantic space, Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, Hissar, Bulgaria, pp. 365–372.
10. *Kolb, P.* (2008), Disco: A multilingual database of distributionally similar words, Proceedings of KONVENS-2008, Berlin.
11. *Mikolov T., Chen K., Corrado G., Dean J.* (2013a), Efficient Estimation of Word Representations in Vector Space, Proceedings of Workshop at International Conference on Learning Representations.
12. *Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J.* (2013b), Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems, pp. 3111–3119.
13. *Mikolov, T., Wen-tau Yih, Zweig, G.* (2013c), Linguistic Regularities in Continuous Space Word Representations, Proceedings of NAACL HLT, pp.746–751.
14. *Panchenko, A., Loukachevitch, N., Ustalov, D., Paperno, D., Meyer, C., Konstantinova, N.* (2015), Russe: The first workshop on Russian semantic similarity, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2015" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2015"], Moscow.
15. *Panchenko, A., Romanov, P., Morozova, O., Naets, H., Philippovich, A., Romanov, A., Fairon, C.* (2013), Serelex: Search and visualization of semantically related words, Advances in Information Retrieval, Springer Verlag, pp. 837–840.

16. *Pekar, V., Staab, S.* (2003), Word classification based on combined measures of distributional and semantic similarity, Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, pp. 147–150.
17. *Pekar, V.* (2004), Distributivnaja model sochetaemostnyh ogranichenij glagolov [A distributional model of verbal selectional restrictions]. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2004" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2004"], Moscow.
18. *Rehurek, R., Sojka, P.* (2010), Software framework for topic modelling with large corpora, Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks, Valletta, Malta, pp. 46–50, 2010.
19. *Sahlgren, M.* (2006), The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces, PhD thesis, University of Stockholm.
20. *Schutze, H.* (1992), Dimensions of meaning, Proceedings of Supercomputing'92, pp. 787–796.
21. *Stefanowitsch, A., Gries, S. T.* (2005), Collostructions: Investigating the interaction of words and constructions, International journal of corpus linguistics, Vol. 8(2), pp. 209–243.
22. *Tian, Z., Xiang, H., Liu, Z., Zheng, Q.* (2013), A Random Walk Approach to Selectional Preferences Based on Preference Ranking and Propagation, Proceedings of the ACL Meeting, 2013, pp. 1169–1179.
23. *Vecchi, E. M., Baroni, M., Zamparelli, R.* (2011), (Linear) maps of the impossible: capturing semantic anomalies in distributional space, Proceedings of the Workshop on Distributional Semantics and Compositionality, pp. 1–9.
24. *Widdows, D., Cohen, T.* (2010), The semantic vectors package: New algorithms and public tools for distributional semantics, IEEE Fourth International Conference on Semantic Computing (ICSC), pp. 9–15.

THE DISCURSIVE CONSTRUCTION *ДЕЛО В ТОМ, ЧТО* AND ITS PARALLELS IN OTHER LANGUAGES: A CONTRASTIVE CORPUS STUDY[1]

Dobrovol'skij D. (dm-dbrv@yandex.ru)
Russian Language Institute, Russian Academy of Sciences, Moscow, Russia

Pöppel L. (ludmila.poppel@slav.su.se)
Stockholm University, Department of Slavic and Baltic languages, Finnish, Dutch and German, Stockholm, Sweden

The primary goal of the present study is to improve methods for contrastive corpus investigations. Our data is the Russian construction *дело в том, что* and its parallels in English, German and Swedish. This construction, which appears to present no difficulty for translation into other languages, is in fact language-specific with respect to at least one parameter. It displays a large number of different parallels (translation equivalents) in other languages, and possesses a complex semantic structure. The configuration of semantic elements comprising the content plane of this construction is unique. The empirical data have been collected from the corpus query system Sketch Engine, subcorpus OPUS2 Russian, and the Russian National Corpus (RNC). The analysis shows that the construction *дело в том, что* has more than 50 parallels in English, over 30 in German, and about 30 in Swedish. In all three languages the most common means of translating the construction is to omit it. Also frequent are the English equivalents *the fact/thing/point/truth is (that)*; *(it's/this/that is) because*; the German expressions *nämlich*; *die Sache ist, die*; *denn*; and the Swedish constructions *saken är den att*; *problemet/faktum är att*. The semantic structure of *дело в том, что* includes the following components: 1) substantiation of something stated previously; 2) indication of the reason something has happened; 3) emphasis on the significance of what has been stated. The different translations of the construction are motivated by the fact that each specific context focuses on one of these meanings.

Keywords: parallel corpus, construction, contrastive corpus analysis, Russian, English, German, Swedish, language specificity, semantics

[1] This study received financial support from the RFFI, grant 16-06-00339.

1. Introduction

At first glance, the construction *дело в том, что* appears to present no difficulty for translation into other languages. It seems to possess complete compositionality, and even its literal translation is confirmed in the parallel corpora. Cf. the following examples from Russian-English (1), Russian-German (2) and Russian-Swedish parallel corpora.[2]

(1) ***Дело в том, что*** *редактор заказал поэту для очередной книжки журнала большую антирелигиозную поэму.* [М. Булгаков. Мастер и Маргарита]

The thing was that *the editor had—commissioned from the poet a long anti-religious poem for the next issue of his journal.*

(2) *Ну, все равно, вы простите меня, — продолжал он, — но* ***дело в том, что*** *это ужасно, ужасно, ужасно!* [Л. Толстой. Крейцерова соната]

Nun, gleichviel, nehmen Sie es mir nicht übel,—fuhr er fort,—aber ***die Sache ist*** *eben* ***die, daß*** *das alles so entsetzlich, entsetzlich, entsetzlich ist!*

(3) ***Дело в том, что****... в последнее время, когда бы я ни пытался... что бы я ни сказал, все ее выводит из себя ... как будто она вообще не хочет слушать. Как будто она орет на меня постоянно. И я не знаю почему.* [Sketch Engine]

Saken är den*… när jag väl pratar blir hon otålig, som om hon egentligen inte ville höra. Det är som om hon är förbannad på mig och jag vet inte varför.*

The English expression *the thing was that,* the German *die Sache ist die, daß* and the Swedish *saken är den* are practically word-for-word translations of the Russian *дело в том, что*.

2. Research goals and data

The goal of the present study is to improve methods for contrastive corpus investigations. The following hypotheses will be tested on the basis of the materials of parallel corpora:

(1) The Russian expression *дело в том, что* has a large number of various parallels in other languages, and the choice of each variant depends on specific contextual conditions.

(2) Despite its apparent simplicity, *дело в том, что* has a complex semantic structure. The configuration of semantic elements comprising the content plane of this construction is unique.

---

[2] Examples (1) and (2) are from the parallel corpora of the RNC, while example (3) comes from the parallel corpus OPUS2 Russian of the query system Sketch Engine.

(3) The expression *дело в том, что* is language-specific with respect to at least one of its parameters.

The empirical data have been collected from the corpus query system Sketch Engine, subcorpus of parallel texts OPUS2 Russian, which contains 307709872 tokens (15.02.2016); the Russian National Corpus (RNC), Russian—English, English—Russian, Russian—German and German Russian subcorpora of parallel texts.

3. Theory and methods

Analysis of the corpus data allows us to identify only the degree of variety in the means of translating a given expression into other languages. When one or another expression lacks a generally accepted standard context-independent translation equivalent, we can speak of an absence of systematic equivalents, i.e., a kind of non-equivalence. Whether such non-equivalence is connected with the category of language-specificity remains an open question.[3] Some interesting thoughts on this subject are presented in [Shmelev 2015], where the following parameters of language-specificity are identified. The first parameter is connected with the number of languages which lack the given phenomenon. The more such languages are discovered, the more possible it is to consider the phenomenon language-specific. The second parameter consists in the specificity of their content aspect (including connotations, background components of meaning, etc.), from which it follows that the degree of distinctiveness of the semantic configuration of an expression is directly proportional to its degree of language-specificity. The third parameter is immediately connected with the second: the more distinctive the semantic configuration of a lexical unit, the more difficult it is to find an adequate translation equivalent of this unit in another language. Here, as Shmelev notes, it must be kept in mind that the object of translation is not individual words but texts, so that the translator can deviate from exact equivalence on the lexical level. Such deviations do not necessarily imply that the translated units are language-specific. Nevertheless, it is natural to interpret the proposal of a large number of different translation equivalents as indicating the absence of a systematic equivalent. This allows us to measure quantitatively the degree of language-specificity in accordance with this third parameter, which is in fact at the focus of the present study.

The method of our analysis is based on considerations presented in [Buntman et al. 2014]. What it essentially entails is determining how many translation equivalents exist for potentially language-specific lexical units and then evaluating their dispersion. The expedience of calculating dispersion as a means of determining the statistical value of the scatter is in need of further substantiation. Other statistical instruments may prove to be more adequate.

The Russian expression *дело в том, что* was submitted to Sketch Engine, and we searched OPUS2 Russian for the construction and its translation equivalents

[3] On language-specificity see [Wierzbicka 1992, 1996; Zaliznyak, Levontina, Shmelev 2005, 2012; Zaliznyak 2015, Shmelev 2002, 2014, 2015].

in English, German and Swedish. Because there were very few parallel texts in all four languages, the search was done for each specific pair of languages: Russian and English, Russian and German, Russian and Swedish. We processed our findings manually to avoid information noise. The search system does not indicate from which or into which language a given context was translated. To ascertain the means by which the Russian construction *дело в том, что* is translated into other languages and the possibilities of this expression appearing in translation from other languages into Russian we used the parallel corpora of the Russian National Corpus (RNC)—English-Russian, Russian-English, German-Russian and Russian-German. Only the corpus query system Sketch Engine contains Swedish-Russian and Russian-Swedish corpora. Let us move on to our findings.

4. Results and discussion

We begin the discussion of results with a presentation of the English results in Sketch Engine. Cf. Table 1.[4]

Table 1. *Дело в том, что*: English parallels Sketch Engine[5]

English parallels	Frequency
zero equivalence	154
the fact is (that)	123
the thing is (that)	98
the point is (that)	70
(it's/this/that is) because/because of	40
it's just (that)/it's that/just/this is that	27
in fact	26
the truth is (that)	26
however	16
the fact of the matter is (that)	15
indeed	13
the problem is (that)	12
you see	9
the reason is (that)	8
as a matter of fact	5
for	5
it's/this is about	5
it happens that/as it happened/what has happened is/what is happening is	5

[4] Single hits are not shown in any of the tables.

[5] The tables do not take into account differences in verb tenses. *The thing is (that)*, for example, includes contexts with all other tense forms.

English parallels	Frequency
the matter is (that)	4
but	4
since	4
it's a fact that	4
well	3
basically	3
what's true is (that)/it was true (that)	3
the consequence is (that)	3
the truth of the matter is (that)	2
the answer is (that)	2
the concern is (that)	2
the crux of the matter is (that)	2
the question is (that)	2
as	2
you know	2
look	2
the position is (that)	2
the thing about	2
in effect	2

Besides the data presented in the table, more than 43 single English correlates of the construction *дело в том, что* were found: *the situation is*; *that means that*; *my story is*; *the issue is*; *the reality is*; *the content is*; *the explanation is*; *the fact remained that*; *the fact that*; *this is due to*; *it has everything to do with; what I'm trying to say is that*; *except that*; *that is*; *in reality*; *actually*; *in practice*; *the word is*; *the plan was; here's the thing*; *this is the situation*; *sort of*; *the point being*; *the purpose of*; *it is not that*; *thus; it should be noted that*; *in truth*; *for the reason that*; *as it was*; *rather*; *in that it is*; *that is*; *instead*; *namely*; *in that connection*; *in this regard*; *it is which*; *to be blunt*; *here too*; *it is a matter of*; *accordingly*; *the trouble is.*

The first thing that stands out here is the large number of different English parallels of the Russian construction. In all we found 80 such equivalents.[6] The second important feature is that of these 80, 43 occur only once, which indicates significant scattering in these English parallels.

Let us now turn to the Russian-English parallel corpus of the RNC. Cf. Table 2.

[6] Some of them have been analyzed in [Aijmer 2007; Delahunty 2011, 2012].

Table 2. *Дело в том, что*: English parallels, RNC Russian—English

Engish parallels	Frequency
zero equivalence	27
the fact is (that)	14
the thing is (that)	14
the point is (that)	10
you see	3
actually	2
in point of fact	2
the matter is (that)	2

In this corpus we found 26 translation equivalents, of which 18 occur only once. It is evident from the table that the results partly coincide and partly diverge. Four of the most frequent equivalents—zero equivalent, *the fact is (that)*; *the thing is (that)* and *the point is (that)*—completely coincide, which indicates that the findings are non-random. At the same time, the relatively frequent constructions found in Sketch Engine—*in fact*; *the truth is (that)* and *however*—do not occur in the RNC, but *(it's/this/that is) because/because of*; *it's just (that)/it's that/just/this is that* and *the fact of the matter is (that)*—occur only once. These divergences are quite natural. Sketch Engine is much larger than the RNC, while the RNC is much cleaner. In addition, in Sketch Engine it is impossible to determine which language is the source language, and the texts in these corpora differ with respect to genre. The RNC contains almost exclusively fictional texts, whereas non-fiction dominates in Sketch Engine.

Now for our analysis of the German materials. Cf. Tables 3 and 4.

Table 3. *Дело в том, что*: German parallels, Sketch Engine

German parallels	Frequency
zero equivalence	19
die Sache ist die (dass)	8
aber	5
es geht darum, dass	4
es ist (doch) so, dass	3
die Wahrheit ist, dass	3
wissen Sie	2
nur (dass)	2
Tatsache ist (nun mal)	2
es ist nur (dass)	2
ich meine	2
der Punkt ist	2

Table 4. *Дело в том, что*: German parallels, the RNC Russian—German

German parallels	Frequency
die Sache ist die (dass)	18
zero equivalence	11
nämlich	9
es handelt sich darum, dass	3
die Hauptsache ist, (dass)	3
doch	2
der Grund (war, dass)	2
es kommt (vielmehr/doch nur) darauf an	2

Although the Russian-German materials are considerably smaller in scope, the results exhibit tendencies similar to those observed in the English materials. In Sketch Engine we found 20 German parallels and in the RNC 13, some of which coincide and some do not. Two of the most frequent parallels in Sketch Engine—zero equivalent and *die Sache ist die (dass)*[7]—coincide with the most frequent ones in the RNC, although in reverse order. The most important difference is the absence of *nämlich* in Sketch Engine, whereas in the RNC it occurs 9 times. This difference is significant because even a superficial analysis of the word *nämlich* shows that its communicative function is very close to that of the Russian construction *дело в том, что*. On the whole, the German parallels in Tables 3 and 4 display considerable scatter. What most convincingly argues that *дело в том, что* is non-equivalent with respect to German is the partial absence of a translation equivalent in the parallel texts.

Table 5 shows the data from Sketch Engine, which is the only corpus of texts at our disposal containing Swedish parallels.

Table 5. *Дело в том, что*: Swedish parallels, Sketch Engine

Swedish parallels	Frequency
zero equivalence	45
saken är den att	16
men	8
problemet är att	7
faktum är att	4
det viktiga är (att)/det är viktigt att	4
det är för att	4
sanningen är att	3
grejen är den att	3
poängen är att	3
för (att)	3
det handlar om att	2

[7] For analysis of *die Sache ist (dass)* cf. [Günthner 2008].

Swedish parallels	Frequency
det vad jag vill säga är att	2
i själva verket	2
jag/han menar att	2

We found 25 Swedish parallels, the most frequent of which—zero equivalent and *saken är den att*—once again testify to a certain non-translatability of the Russian expression *дело в том, что*. In the intermediate zone (from 2 to 13) there are 13 equivalents, 10 equivalents are used only once. Here as well we can speak of considerable scatter; that is, relative to Swedish the construction *дело в том, что* is difficult to translate.

Shmelev [2015] points out that for determining the language-specificity on the basis of the parameter of non-equivalence, the parallel corpora in which the language of the expression under analysis is the target rather than the source language have the best diagnostic potential. The appearance of such words is evidently the unconscious solution of the translator as a native speaker. For that reason we turned to the English-Russian and German-Russian parallel corpora of the RNC. Cf. Tables 6 and 7.

Table 6. *Дело в том, что*: English—Russian, the RNC

English parallels	Frequency
zero equivalence	38
the fact is (that)	36
for	34
it’s just (that)/it’s that/just/this is that	16
(that is) because	14
(as) you see	11
well	7
the thing is (that)	7
but	5
it happens (that)	4
actually	4
the truth is (that)	4
the point is (that)	4
in fact	4
the reason is (that)	3
the problem is (that)	2
I mean	2
as a matter of fact	2
I tell you	2
in truth	2
to begin with	2

The English-Russian corpus shows even greater scatter among the English equivalents of *дело в том, что*: in all, there are 54 different translations, of which 6 equivalents occur more than 10 times each, 15 are found in the range from 2 to 10 times, and 33 occur only once. It is reasonable to compare these findings with those from the Russian-English parallel corpus of the RNC (Table 2), since they are basically comparable. The corresponding results from this corpus are as follows: 3 are used more than 10 times each, 5 in the range from 2 to 10 times, and 18 occur only once. Of the most frequent equivalents, only two—zero equivalent and *the fact is (that)*—coincide. Our assumption that translators from Russian to English most often follow the form of the original, using constructions such as *fact is (that)*; *the thing is (that)* and *the point is (that)*, is fully corroborated. Translators from English to Russian, on the other hand, more often employ the language-specific expression whenever it is not dictated by form. Thus the most frequent group in Table 6 includes lexical units such as *for, just, because, you see*. In addition, syntactic means such as cleft sentences are also used. Cf. (4).

(4) *They didn't believe me at first either, he said.* ***It's just that*** *we get a lot of calls like this. But I believe you, Doctor, so why don't you continue with the story?*
[M. Connelly. City of Bones]

*Они тоже мне сначала не верили,—промолвил Гийо.—****Дело в том, что*** *мы получаем много подобных звонков. Но я верю вам, доктор, и, прошу, продолжайте.*

A comparison of the German-Russian and Russian-German parallel corpora of the RNC produces very similar findings. Cf. Table 7 and Table 4 above.

Table 7. *Дело в том, что*: German—Russian RNC

German parallels	Frequency
nämlich	27
zero equivalence	11
die Sache ist die, (dass)	10
denn	8
eben	3
aber	3
es kommt darauf an	2

The following features stand out. The formal correlate *die Sache ist die, (dass)* dominates in translations from Russian to German, while in the German-Russian corpus the word *nämlich* often correlates with *дело в том, что* although their formal structures have nothing in common. This confirms what was stated earlier. Cf. (5).

(5) *Wir landeten, und—die ganze Insel bestand aus einem großen Käse. Wir hätten dies vielleicht gar nicht entdeckt, wenn uns nicht ein sonderbarer Umstand auf die Spur geholfen hätte. Es war* ***nämlich*** *auf unserm Schiffe ein Matrose, der eine*

natürliche Antipathie gegen den Käse hatte. [G. A. Bürger. Die Abenteuer des Freiherrn von Münchhausen]

Как выяснилось, весь остров представлял собой большой сыр. Возможно, мы даже не заметили бы этого, если бы не одно обстоятельство, открывшее нам истину. ***Дело в том, что*** *у нас на корабле находился матрос, отроду испытывавший отвращение к сыру.*

Another feature of the German-Russian corpus is that the group of relatively frequent parallels includes the causal word *denn*, which is similar to the English conjunctions because and for in the English-Russian corpus.

5. Conclusion

The study advanced three hypotheses, each of which may be considered confirmed. The data presented in the investigation show that the construction *дело в том, что* has many different English, German and Swedish translation equivalents. It is important to note that most of these equivalents are not synonymous among themselves, which means that the choice of one or another of them depends not on the subjective preferences of the translator, but on contextual factors. Following from this as well is the fact that the construction *дело в том, что* has a complex semantic configuration that is unique relative to English, German and Swedish. The semantic structure of *дело в том, что* includes at least the following meanings: 1) substantiation of something stated previously; 2) indication of the reason something has happened; 3) emphasis on the special significance of the following clause. Equivalents from the various groups are selected depending on which meaning is being highlighted in the utterance. This can be illustrated with the following examples from Sketch Engine and the RNC.

1) English *you see* in ***Дело в том, что*** *это ваша мать—****You see****, it’s your mother*), German *nämlich* in ***Дело в том, что*** *она была значительно старше меня—Sie war* ***nämlich*** *bedeutend älter als ich* or Swedish *det var så att* in ***Det var så att*** *jag råkade kalla henne subba—****Дело в том, что*** *я случайно назвал ее кошелкой.* Russian *дело в том, что* not only substantiates previous statements but also indicates the reason in each statement and emphasizes the significance of the following clause.
2) Causal conjunctions and constructions such as English (*it is*) *because, the reason is that*, German *denn*, and Swedish *för.* In the following examples English *it is because*, Swedish *för* and German *denn* indicate the reason for something that has been stated in the previous utterance: *То есть,* ***дело в том, что*** *я — женщина!—So* ***it is because*** *I’m a woman*; *Отличная книга. Её должен прочесть каждый. Видишь ли,* ***дело в том, что*** *если мы будем беспечны…—Gutes Buch. Jeder sollte es lesen.* ***Denn****, weißt du, es geht darum, wenn wir nicht aufpassen…*; *Может быть, вы поможете мне…* ***дело в том, что*** *мне очень нужно с ним встретиться. Du kanske kan hjälpa mig,* ***för*** *det är viktigt att jag får tag i honom.*

3) Focusing particles and constructions such as English *the point is, the thing is*, German *eben, die Sache ist die, (dass)* or Swedish *saken är den att, poängen är den att*. English *the point is that in* ***Дело в том, что*** *этот противник никогда не отступает—**The point is that** this enemy never retreats*; German *was die Hauptsache ist* in *Но* ***дело в том, что*** *Анну я вам не отдам—**Aber was die Hauptsache ist**: ich lasse auf Anna nichts kommen*; and Swedish *poängen är att* in *Нет, но* ***дело в том, что*** *искушение всегда остается—**Nej, men poängen är att** frestelsen alltid är där* focus on the importance of what is said. Russian *дело в том, что* does not only focus the following clause but conveys other meanings as well—substantiation and indication of the reason.

The uniqueness of the semantic and conceptual configuration determining the meaning of the construction testifies to its language-specificity relative to the languages examined here.

The findings of this and similar studies can be useful not only in developing the theory of language specificity but also in lexicography, translation and pedagogy. What we consider to be our primary accomplishment is that we have succeeded in outlining a new approach to working with parallel corpora in contrastive corpus investigations.

References

1. *Aijmer K.* (2007), The interface between discourse and grammar: *The fact is that*, in Connectives as discourse landmarks, Benjamins, Amsterdam, Philadelphia, pp. 31–46.
2. *Buntman N. V., Zaliznyak Anna A., Zatsman I. M., Kruzhkov M. G., Loshchilova E. Yu., Sitchinava D. V.* (2014), Informational technology in corpus-based studies: towards a cross-linguistic database, Informatics and its applications [Informacionnye tekhnologii korpusnykh issledovaniy: printsipy postroeniya krosslingvisticheskikh baz dannykh], Informatics and its applications [Informatika i ee primeneniya], vol. 8, issue 2, pp. 98–110.
3. *Delahunty G. P.* (2011), Contextually determined fixity and flexibility in *thing* sentence matrixes, Yearbook of Phraseology 2, Mouton de Gruyter, Berlin, N. Y., pp. 109–135.
4. *Delahunty G. P.* (2012), An analysis of *the thing is that S* sentences, Pragmatics, 22 (1), pp. 41–78.
5. *Dobrovol'skij D., Pöppel L.* (2015a), Corpus perspectives on Russian discursive units: semantics, pragmatics and contrastive analysis, Yearbook of corpus linguistics and pragmatics, Springer International Publishing, pp. 223–242.
6. *Dobrovol'skij D., Pöppel L.* (2015b), Russian constructions *то-то и N* and *в том-то и N* and their English and Swedish equivalents: a corpus-based cross-linguistic analysis, in Trends in Slavic Studies, URSS, Moscow, pp. 595–607.
7. *Dobrovol'skij D., Pöppel L.* (2015c), Entrenched lexical patterns: the Russian construction *в том-то и весь N,* Procedia—Social and Behavioral Sciences 206, Elsevier, pp. 18–23.

8. *Günthner S.* (2008), ”Die *die Sache/das Ding ist*-Konstruktion im gesprochenen Deutsch—eine interaktionale Perspektive auf Konstruktionen im Gebrauch, in Konstruktionsgrammatik II. Von der Konstruktion zur Grammatik, Stauffenburg, Tübingen, 157–177.
9. *Shmelev A. D.* (2015), Russian language-specific lexical units in parallel corpora: prospects of investigation and “pitfalls” [Klyuchevye slova: perevod, parallel’nuy korpus, leksicheskaya edinica, semanticheskoe razlichie, lingvospecifichnost’, “neperevodimost’”], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2015” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2015”], Moscow, pp. 584–595.
10. *Wierzbicka A.* (1992). Semantics, culture, and cognition. Universal human concepts in culture-specific configurations, Oxford Univ. Press, N. Y., Oxford.
11. *Wierzbicka A.* (1996). Semantics: primes and universals, Oxford Univ. Press, Oxford.
12. *Zaliznyak Anna A.* (2015), Russian language-specific words as an object of contrastive corpus analysis [Lingvospecifichnye edinicy russkogo jazyka v svete kontrastivnogo korpusnogo analiza], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2015” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2015”], Moscow, pp. 65–662.
13. *Zaliznyak Anna A., Levontina I. B., Shmelev A. D.* (2005), Key ideas of the Russian language picture of the world [Klyuchevye idei russkoy yazykovoy kartiny mira], Yazyki slavyanskoy kul’tury, Moscow.
14. *Zaliznyak Anna A., Levontina I. B., Shmelev A. D.* (2012), Constants and variables of the Russian language picture of the world [Konstanty i peremennye russkoy yazykovoy kartiny mira], Yazyki slavyanskoy kul’tury, Moscow.

AUTOMATIC GENERATION OF THE DOMAIN-SPECIFIC SENTIMENT RUSSIAN DICTIONARIES

Dubatovka A. (alina.dubatovka@gmail.com)
Saint Petersburg State University, St. Petersburg, Russia

Kurochkin Yu. (yurakura@yandex-team.ru)
Yandex, St. Petersburg, Russia

Mikhailova E. (e.mikhaylova@spbu.ru)
Saint Petersburg State University, St. Petersburg, Russia

This paper presents an algorithm for generating the Domain-Specific Sentiment Russian dictionary using a graph model. It is important to emphasize that the described algorithm does not require any human-labeling, but just a sufficiently large corpus of Russian texts from the subject area, which can be generated automatically for most domains. Our algorithm is not strictly confined to the Russian language and, if necessary, can be generalized to develop dictionaries in other languages.
Dictionaries of positive and negative words are created using the analysis of the graph constructed on unlabeled corpus of the Domain-Specific Russian texts. The graph was built using the approach described in [6], pre-adapted to texts in Russian. The applicability of this method to create a graph for prediction of polarity of adjectives in reviews in Russian language is experimentally evaluated.
The original method of graph processing for splitting the vertex set of this graph into subsets of positive and negative words was proposed and implemented. The algorithm starts with gathering a small seed set of adjectives, polarity of which is unambiguous irrespective of a subject area (for example, “bad”, “good”, “terrible”, “excellent”).
Further, words are distributed iteratively: each time a vertex is added to the set, if the vertex is most strongly associated with the already existing vertices in the set. Several weighting functions on the edges were compared, as well as functions of attraction to the sets of positive and negative words with the aim of composing the most accurate dictionaries of positive and negative adjectives for a specific subject area.

Keywords: sentiment analysis, sentiment lexicon, opinion mining

1. Introduction

Public opinion and review studies have become an important part of decision-making processes. When a particular choice is associated with financial expenses (purchase of any goods, services, etc.), customers often rely on other people’s experience.

Information obtained from such studies is one of the most important factors in the final choice made.

Recently, the task of "opinion mining" has aroused significant interest among researchers of natural language texts as well as has become an important constituent of a decision-making process. Automatic sentiment analysis of a text finds a broad application in different fields of human activity—politics, business, film industry. Information obtained from review analysis affects the final choice that people make. Due to the broad variety of application areas, the actual goal is not just the review analysis itself, but also the opinion extraction from domain-specific texts. Since most of the algorithms for the review analysis are based on the use of sentiment dictionaries—dictionaries of positive and negative meaning of words, the automatic generation of such dictionaries becomes an important task. Solution of such kind of tasks strongly depends on the domain-specific area and language features, as the same words used in different situations can give diametrically opposite polarity. For example, the word "thin" is positive for laptop characteristics, while it would be uncomfortable to stay in a hotel with "thin walls". There are various situations when different meanings of one and the same word could be associated with alternative sentiments.

This paper describes an algorithm for generating the Russian sentiment dictionary for a domain-specific area using a graph model. We investigate the dependence of the dictionary's quality on graph building and analysis parameters. It is important to note that the described algorithm does not require any preliminary marking, but only a sufficiently large corpus of Russian texts from the domain area, which can be easily prepared by the user him-/herself. The algorithm is not bound to the Russian language and, if desired, can be generalized to create dictionaries of other languages.

Dictionaries of positive and negative words are created using the analysis of graphs constructed on a raw corpus of Russian texts from the domain area. Such corpus can be generated automatically for most domains. Graph nodes are adjectives, and edges connect the adjectives joined by a coordinating conjunction at least in one sentence.

The process of splitting graph nodes into positive and negative word subsets starts from small initial sets consisting of adjectives, the sentiment polarity of which is unambiguous irrespective of the subject area (*e.g*, "*bad*", "*good*", "*terrible*", "*excellent*"). Further, the remaining words are distributed iteratively: each time a node is added into the set most strongly associated with the already existing nodes. We compared efficiency of several weighting functions on the edges as well as distance functions to compile the most accurate dictionaries. This work has also demonstrated applicability of the approach described in [6] for sentiment analysis of adjectives in Russian-speaking reviews.

2. Related work

Over the last five years, there has been a tremendous increase in demand for sentiment analysis tools by companies for monitoring people's opinions on company's products and services and by social science researchers. All sentiment analysis tools rely on dictionaries of words and phrases with positive and negative connotations. Such dictionaries are necessary for different languages and different domain-specific areas affecting

the polarity of words. Thus, the task of building sentiment dictionaries for a particular language and the domain-specific area is highly relevant, because even if there is a large amount of publicly available labeled data, most of such dictionaries are composed for the English language and examine either movie reviews or reviews on equipment.

For example, [1] describes an algorithm for compiling one of few publicly available Russian-language dictionaries of opinion words for the product meta-domain with the help of learning several classifiers on one domain and then migrating the resulting model to other domain areas. Further, in [7], authors try to clarify the dictionary obtained by analysis of the corresponding subgraph of the RuThes thesaurus.

The task of creating the sentiment vocabularies is relevant not only for Russian language but also for many other languages. Thus, in [10] authors propose a method for automatic dictionary creation for new languages (German, Russian, Italian, French, Arabic and Czech) using manually compiled dictionaries in English and Spanish.

In [2], the original manually compiled dictionary for the German language was enlarged by the construction of the graph based on the untagged German corpus, as described in [6], and by its further analysis using the classification method of maximum entropy.

Different graph models are widely used for subtasks such as adapting the model to a new domain area, highlighting sentiment sentences from the text or ranking words according to the opinion polarity. The authors of [11], having the corpus from marked up documents in one domain and unlabeled corpus from a different domain, determine the polarity by building and analyzing a weighted graph composed with the feedback as nodes and the cosine measure of similarity between documents as weight of the edge. In [8] it is proposed to build a graph using sentences and relationships between them. The problem of automatic detection of sentiment sentences from the text is solved by searching the minimum cut in the graph.

Since graph models play a key role in the social network analysis, various algorithms have been developed for their analysis. Some of them could be applied to the problem considered in this paper. For example, in [3, 4] the various algorithms of random walks (in particular, PageRank) are adapted to the graph constructed on the basis of eXtended WordNet [5] to rank the sentiment polarity of words.

3. Methodology

As shown in [6], coordinating conjunctions, connecting coordinate adjectives and adverbs convey the relation of the polarity of the connected words. As a rule, copulative conjunctions are placed between words that have the same polarity («*Tasty and healthy Breakfast*"), and the adversative conjunctions are placed between words with nearly opposite sentiment ("*Cheap but nice hotel*").

These relations between the word sentiments allow to build a weighted graph whose nodes are the adjectives and edges are connections between them, labeled with the number of sentences with the words connected either by the copulative or adversative conjunction.

Analysis of the obtained graph permits to evaluate the "positivity" or "negativity" of words, which are its nodes: the better the node is connected with other "positive" nodes and the worse with the "negative", the more positive it is.

Thus, the algorithm for constructing the sentiment dictionary consists of two main stages, described below: building the graph of connections between words and its processing.

3.1. Constructing the graph

As described above, we construct the graph using adjectives as nodes. The edges are copulative and adversative relations between them. To build such edges, we extract coordinate adjectives from the texts and consider connections between them. The adjectives are coordinate if they are consistent in their gender, number and case, and satisfy the template

$$(\text{ADV} \mid \text{NEG}) * \text{ADJ}(, ? (\text{AND} \mid \text{BUT})? (\text{ADV} \mid \text{NEG}) * \text{ADJ}) +,$$

where AND is the conjunction "and", BUT is one of adversative conjunctions ("but", "instead", "however", "nevertheless "), NEG is a negation, ADV is an adverb of measure and degree ("very", "quite", "too", "completely") and ADJ is an adjective.

The sentiment link was formed for each pair of the selected coordinate adjectives (either positive or negative depending on the conjunction). This link is necessary to calculate the weight of an edge. For example, for the phrase "*Tasty, plentiful but not very varied and expensive breakfast*" three positive links are produced: (*tasty, plentiful*), (*tasty, varied*), (*plentiful and varied*); and three negative links are: (*tasty, expensive*), (*plentiful, expensive*), (*varied, expensive*).

To determine the part of speech and word forms we use the morphological analyzer of the Russian language Mystem[1] [9]. Mystem works on the basis of the dictionary and normalizes words into the primary form, and also processes their grammatical information. For words missing in the dictionary, Mystem performs a hypothetical analysis of words.

In case the negation comes before the adjective, an orientation of connection between words reversed. For example, in the sentence "The pool is large, but not very deep", the adjectives "large" and "deep" will have a positive connection, meaning the sentiment coincidence, although they are connected by the adversative conjunction «but». Similarly, by processing the phrase "Delicious and not expensive food", "delicious" and "expensive" will have a negative connection and therefore obtain opposite polarity.

Since we analyzed the feedback of Internet users, not literary texts, it is necessary to consider not only the punctuation rules of the Russian language, but also the most common (though erroneous) forms. So, for example, people sometimes miss a comma, even if grammar rules require to use it: a comma between coordinate adjectives, a comma before "but" or between repeated conjunctions "and", so the template should not be very strict.

---

[1] https://tech.yandex.ru/mystem/

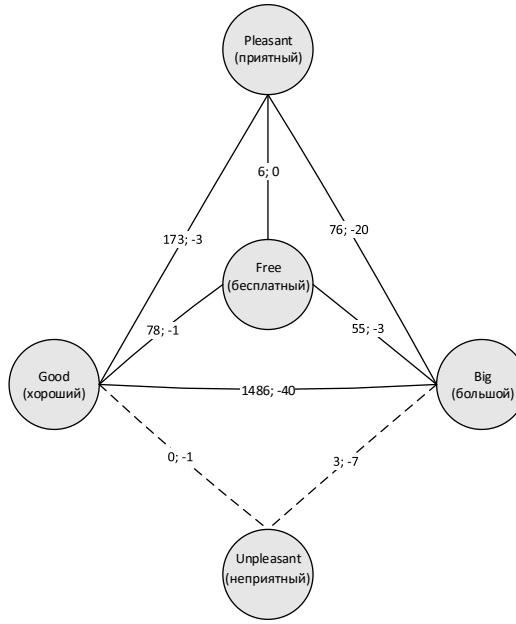


Fig. 1. A fragment of the graph without removal the "un-" prefixes

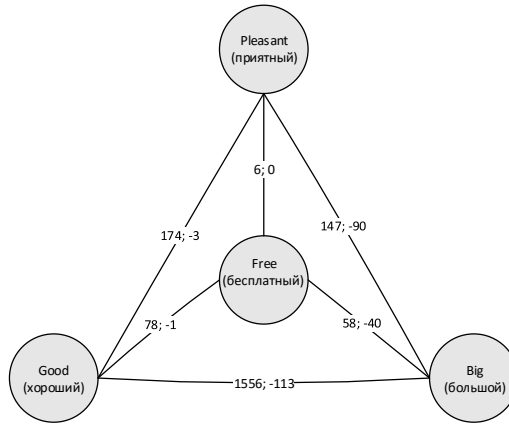


Fig. 2. A fragment of the graph with removal the "un-" prefixes

3.1.1. Particle "not" and the prefix "un-"

Apart from the fact that negation can be expressed by a free-standing particle "not", it can appear in the prefix form "*un-*" with adjective. For example, "*not very pleasant*", "*unpleasant*" and even "*not pleasant*", by and large, represent the same negation of the adjective "*pleasant*". So we analyzed how the negative prefix "*un-*" affects the sentiment adjective. For this purpose, we have implemented and compared two

approaches. The first approach is to consider such adjectives (for example "*unpleasant*" and "*pleasant*") as two different nodes of the graph. The second one is to separate the prefix "*un-*" if it is possible (i.e. the word without the prefix is identifiable by Mystem) and consider it as a negative particle "not", that is, the adjective "*unpleasant*" is equated with the phrase "*not pleasant*" and, hence, we do not have to create a special node for the word "*unpleasant*". Fig. 1 and 2 present fragments of the graphs built without and with removal the "un-" prefix respectively.

3.2. Graph processing

The next stage is to split the previously obtained graph into two clusters: the positive set and the negative one. To do this, we initialize the positive and the negative sets and iteratively add one by one the non-assigned nodes to each of these sets. The candidate node is added into the nearest set ("positive" or "negative"). The candidates are the nodes with at least one edge connected with these sets. After adding a node into the set, all its neighbors that have not yet been assigned are added to the set of candidates and the distances between the candidate node and the final sets are recalculated.

3.3. Initialization

It is possible to initialize the positive and negative sets according to the fact that the sentiment polarity of certain words is obvious regardless of the context and domain area.

Therefore, the initial set of "*positive*" words contains such apparently positive adjectives like "*good*", "*excellent*", "*wonderful*", "*lovely*", "*best*", but "*negative*" words set consist of negative adjectives "*bad*", "*awful*", "*disgusting*", "*worst*", "*poor*".

3.3.1. Weight of the graph edges

In the course of the graph construction, a pair of numbers is assigned to every edge—the number of positive and negative connections. The question is how to calculate the weight of edges using these numbers. The weight of edges can be calculated using these numbers as a basic difference, as arbitrary linear combination or a nonlinear function. In our experiments, we calculate the edge weight according to the formula

$$weight(word_1, word_2) = \#(word_1 AND\ word_2) - K * \#(word_1 BUT\ word_2),$$

where K is the coefficient of the adversative conjunctions relevance (the number of negative connections is much less than the number of positive connections, so we give greater weight to the first ones).

3.3.2. Distance to the final set

A similar problem arises—when calculating the distance between the candidate node and each of the final sets. Multiple edges of different weights can connect the node with the set. Is one "heavy" edge better than a lot of "light" edges? The node may

have edges connected with the opposite set. We compared the following most common intuitive techniques for distance calculation.

3.3.3. The heaviest edge

The distance between the node and the set is the weight of the heaviest edge connecting each other.

3.3.4. The sum of the weights of edges

The edge weight is the sum of the edge weights connecting with the considered set, subtracted by the sum of weights of the edges connected with the opposite set.

4. Description of experiments

The algorithms' input was supplied with 259,023 depersonalized unlabeled hotel reviews. The size of the dataset was 660 Mb. The reviews were about different hotels over the world. As long as the texts were written by real users, they contained a lot of misspellings and grammatical errors and informal words. As a rule, these reviews described hotel location, rooms, staff, meal and beaches, but a lot of texts contained unrelated information concerning flight, excursions, places of interest *etc*.

In order to evaluate precision of the proposed algorithms on all the data available we have manually labeled all the adjectives extracted by the algorithms into three classes: positive, negative, and neutral. So we obtained "large" dictionaries of positive, negative, and neutral words consisting of 970, 1,000 and 2,591 words respectively. These dictionaries were used for the precision evaluation, because after processing by the algorithm each word resulted in being placed into one of the "large" dictionaries. In this case we calculated not only classical precision, but also a precision of separation positive words from negative ones. For this propose, we discarded all words contained in the "large" neutral dictionary from the result, since the detection of neutral words is actually a separate challenge [1, 8], and then calculated a classical precision. Table 1 contains sizes of "large" dictionaries as well as the resulting dictionaries for the algorithm with and without removing the "*un-*" prefixes.

Table 1. Sizes of the "large" dictionaries and result dictionaries

	Positive	Negative	Neutral	Total
Algorithm without removing the "un-" prefix	5,252	2,815	—	8,067
Algorithm after removing the "un-" prefix	4,936	2,695	—	7,631
"Large" dictionary	1,948	1,946	4,951	8,845

Because of the large amount of input data, a human assessment is impossible, so recall was estimated by using "manual" dictionaries. For this purpose, we manually labeled 500 random reviews from the input data. Every occurring adjective was labeled as positive or negative depending on its sentiment polarity in the review. Thus

we compiled "manual" dictionaries of positive and negative words, consisting of 173 and 127 words respectively, for recall estimation. Since these 500 reviews were selected randomly, and adjective distribution over the reviews is considered to be uniform, we can assume the sample as unbiased, and therefore, the recall calculated for these 500 reviews is a good approximation for the recall of all data. To estimate recall for positive dictionary we calculated what part of words from "manual" positive dictionary occurs in the positive dictionary generated by the algorithm, the same was done for the negative dictionaries. Table 2 contains sizes of "manual" dictionaries and sizes of the intersections of "manual" and result dictionaries for both algorithms.

Table 2. Sizes of the "manual" dictionaries and "small" result dictionaries obtained as the intersection of the result dictionaries and corresponding "manual" dictionaries

	Positive dictionary	Negative dictionary	Total
"Manual" dictionary	173	127	300
Algorithm without "un-" prefix removing	164	74	238
Algorithm with "un-" prefix removing	163	83	246

To study the rate of dictionary degradation and relationship between the result quality and the stop point we built a plot of Precision@n, where Precision@n is a precision of the top n words from each dictionary.

In addition, to explore the dependence of the dictionary quality on the importance coefficient of negative edges K we built a plot of the F_1-measure for different values of parameter K.

5. Results

Tables 3 and 4 contain the results of the algorithms without and with removing the «un-» prefix respectively.

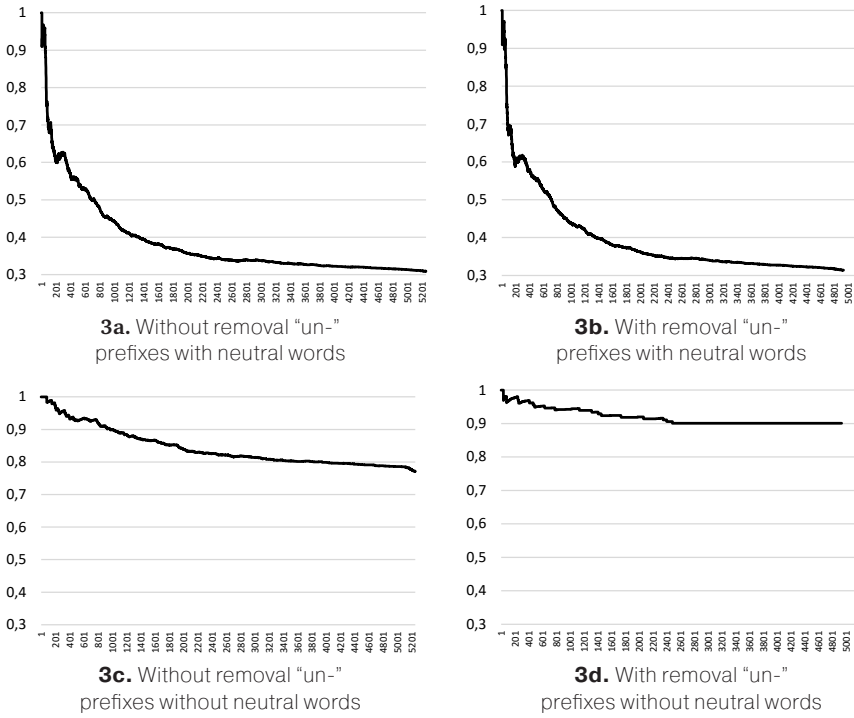
Table 3. The result of the algorithm without removing the "un-" prefix

Metric	Positive dictionary	Negative dictionary	Total dictionary
Recall	0.806	0.684	0.754
Precision	0.309	0.521	0.381
Precision without neutral words	0.770	0.827	0.796
F_1-measure	0.447	0.591	0.506
F_1-measure without neutral words	0.788	0.749	0.774

Table 4. The results of the algorithm after removing the "un-" prefix

Metric	Positive dictionary	Negative dictionary	Total dictionary
Recall	0.793	0.683	0.746
Precision	0.314	0.502	0.380
Precision without neutral words	0.779	0.820	0.799
F_1-measure	0.450	0.579	0.504
F_1-measure without neutral words	0.786	0.745	0.772

Fig. 3 and 4 present plots of Precision@n for positive and negative dictionaries, obtained as a result of processing all reviews, respectively. It is easy to see that the dictionaries start degrading very quickly due to the inclusion of neutral words, however, degradation of the filtered dictionaries, containing sentiment words only, is much slower. We should notice that removing the "un-" prefix does not affect the plot behavior much, if neutral words are taken into account, while for the filtered dictionaries it gives a significant increase in precision especially for the positive dictionary.



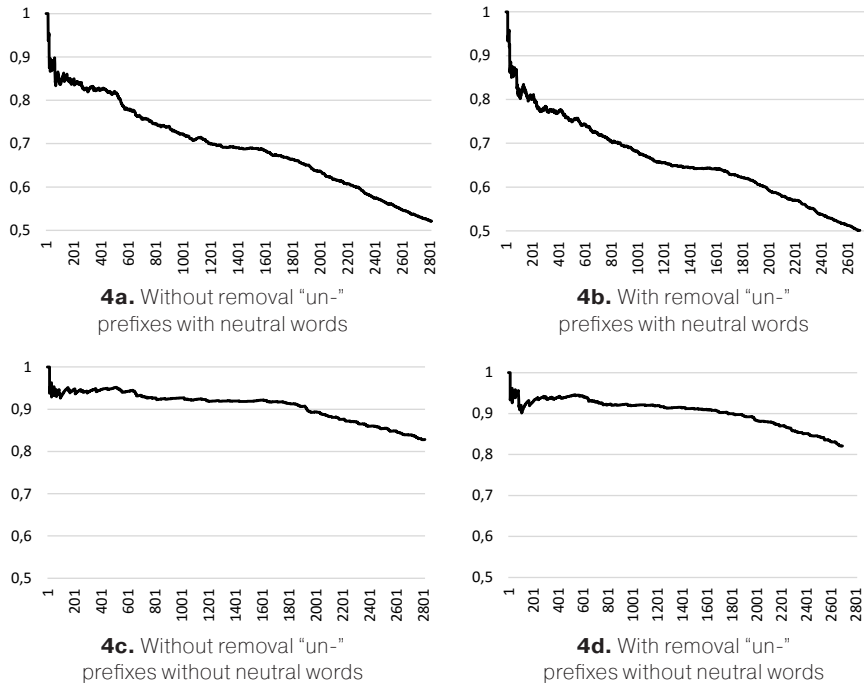
3a. Without removal "un-" prefixes with neutral words

3b. With removal "un-" prefixes with neutral words

3c. Without removal "un-" prefixes without neutral words

3d. With removal "un-" prefixes without neutral words

Fig. 3. Plot of Precision@n for the positive dictionary



4a. Without removal "un-" prefixes with neutral words

4b. With removal "un-" prefixes with neutral words

4c. Without removal "un-" prefixes without neutral words

4d. With removal "un-" prefixes without neutral words

Fig. 4. Plot of Precision@n for negative dictionary

Fig. 5 and 6 present the plots of dependence of F_1-measure with and without neutral words on the parameter K for positive and negative dictionaries. Fig. 7 and 8 contain a scatter plot of recall (the x-axis) and precision (y-axis) with different values of K from 1 to 10 (the points are marked with corresponding parameter values).

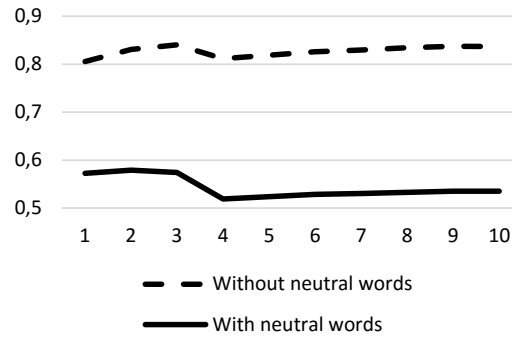


Fig. 5. The dependence of the F_1-measures on the parameter K without removal of "un-" prefix

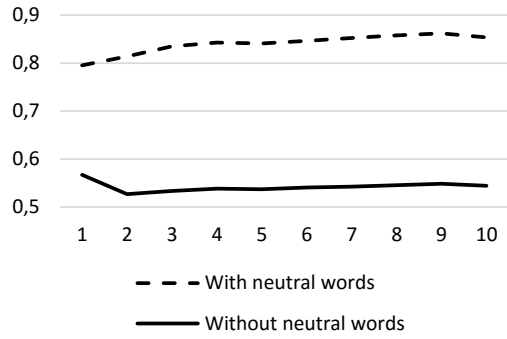


Fig. 6. The dependence of the F1-measures on the parameter K after removing the "un-" prefix

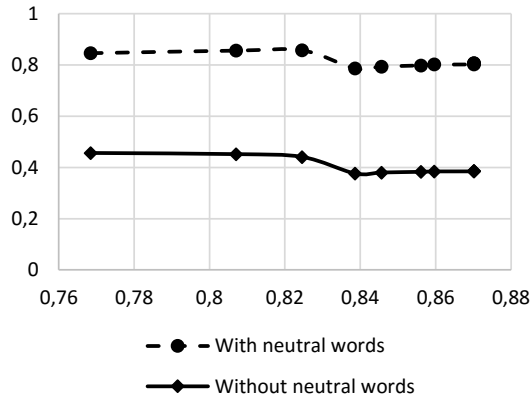


Fig. 7. The values of precision/recall for different values of the parameter K without removal of "un-" prefixes

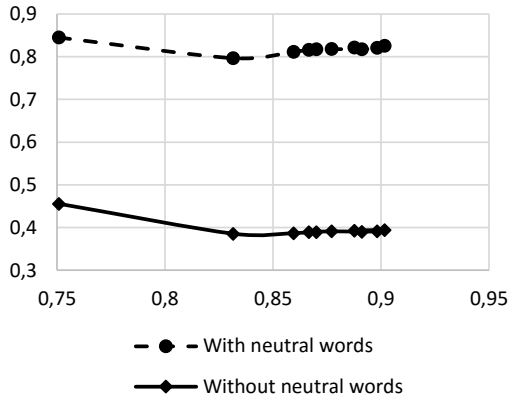


Fig. 8. The values of precision/recall for different values of parameter K after removing the "un-" prefixes

6. Conclusion

This work describes an unsupervised method for building dictionaries of sentiment adjectives based on the analysis of unlabeled reviews from chosen domain. For this propose, we consider and analyze a graph, built using adjectives from the text as the nodes and the syntactic relations between them as edges. To separate the adjectives into positive and negative sets, this graph is split into two clusters using the initial set of «universal» adjectives, whose sentiment polarity does not depend on the chosen domain or context. The paper provides a comparison of several implementations of algorithms for graph constructing and analyzing. These algorithms ensure the construction of dictionaries with 79.9% precision and 75.4% recall.

We considered hotel reviews in Russian language as data for our experiments. The described method is applied to unlabeled texts, and thus input corpus for the algorithm can be formed automatically (e.g., using a crawler), without human assessment. This allows to use this algorithm for an arbitrary domain. Furthermore, this approach can be applied to texts in other languages, as its implementation requires only a morphological analyzer, a list of copulative and adversative conjunctions, and initial set of "universal" sentiment words.

References

1. *Chetviorkin I. I. and Loukachevitch N. V.* (2012), Extraction of Russian sentiment lexicon for product meta-domain, Proceedings of COLING 2012: Technical Papers, Mumbai, pp. 593–610.
2. *Clematide S., Klenner M.* (2010), Evaluation and extension of a polarity lexicon for German, Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), Lisbon, pp. 7–13.
3. *Esuli A., Sebastiani F.* (2007), PageRanking wordnet synsets: An application to opinion mining, ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, pp. 424–431.
4. *Esuli A., Sebastiani F.* (2007), Random-walk models of term semantics: An application to opinion-related properties, Proceedings of LTC 2007, Poznań, pp. 221–225.
5. *Harabagiu S. M., Miller G. A., Moldovan D. I.* (1999), WordNet 2—a morphologically and semantically enhanced resource, Proceedings of SIGLEX99: Standardizing Lexical Resources, College Park, pp. 1–8.
6. *Hatzivassiloglou V., McKeown K.* (1997), Predicting the semantic orientation of adjectives, Proceeding EACL '97 Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, Madrid, pp. 174–181.
7. *Loukachevitch N. V., Chetviorkin I. I.* (2014) Refinement of Russian sentiment lexicons using RuThes thesaurus), Selected Papers of XVI AllRussian Scientific Conference "Digital libraries: Advanced Methods and Technologies, Digital Collections", Dubna, pp. 61–65.

8. *Pang B., Lee L.* (2004), A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, pp. 271–278.
9. *Segalovich I.* (2003), A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine, MLMTA, Las Vegas, pp. 273–280.
10. *Steinberger J., Ebrahim M., Ehrmann M., Hurriyetoglu A., Kabadjov M. A., Lenkova P., Steinberger R., Tanev H., Vazquez S., Zavarella V.* (2012), Creating sentiment dictionaries via triangulation, Decision Support Systems, Vol. 53(4), pp. 689–694.
11. *Wu Q., Tan S., Cheng X.* (2009), Graph ranking for sentiment transfer, ACL/IJCNLP (Short Papers), pp. 317–320.

ВРЕМЕННАЯ КООРДИНАЦИЯ МЕЖДУ ЖЕСТОВЫМИ И РЕЧЕВЫМИ ЕДИНИЦАМИ В МУЛЬТИМОДАЛЬНОЙ КОММУНИКАЦИИ[1]

Федорова О. В. (olga.fedorova@msu.ru)[1,2,3]
Кибрик А. А. (aakibrik@gmail.com)[1,2]
Коротаев Н. А. (n_korotaev@hotmail.com)[2,3,4]
Литвиненко А. О. (allal1978@gmail.com)[2]
Николаева Ю. В. (julianikk@gmail.com)[1,2]

[1]МГУ имени М. В. Ломоносова, Москва, Россия
[2]Институт языкознания РАН, Москва, Россия
[3]РАНХиГС, Москва, Россия
[4]РГГУ, Москва, Россия

Доклад развивает проблематику мультимодальной лингвистики, рассматривающей все каналы передачи информации — вербальные единицы, просодию, жесты, мимику, направление взора и т. д. — в их взаимодействии. Одним из ключевых вопросов мультимодальных исследований является вопрос о временно́й координации между иллюстративными мануальными жестами (спонтанными жестами рук, сопровождающими речь) и элементарными дискурсивными единицами (базовыми единицами локальной структуры дискурса). В настоящей работе этот вопрос рассматривается на материале создаваемого мультимодального корпуса «Рассказы и разговоры о грушах». В ряде исследований было показано, что время начала жеста обычно опережает время начала речи. Для проверки этого положения был разработан аналитический аппарат, который позволил провести более детальное исследование. В результате выяснилось, что в исследованном материале жесты опережают речь менее чем в половине случаев. Наиболее правдоподобное объяснение полученных отличий от работ предшественников связано с функциональными классами жестов, поскольку распределение жестов по классам существенно зависит от жанра дискурса и индивидуальных особенностей говорящих.

Ключевые слова: мультимодальный дискурс, жест, элементарная дискурсивная единица, временная координация

[1] Работа выполнена при финансовой поддержке РНФ (проект № 14-18-03819).

TEMPORAL COORDINATION BETWEEN GESTURAL AND SPEECH UNITS IN MULTIMODAL COMMUNICATION

Fedorova O. V. (olga.fedorova@msu.ru)[1,2,3]
Kibrik A. A. (aakibrik@gmail.com)[1,2]
Korotaev N. A. (n_korotaev@hotmail.com)[2,3,4]
Litvinenko A. O. (allal1978@gmail.com)[2]
Nikolaeva Ju. V. (julianikk@gmail.com)[1,2]

[1]Lomonosov Moscow State University, Moscow, Russia
[2]Institute of Linguistics RAS, Moscow, Russia
[3]RANEPA, Moscow, Russia
[4]RSUH, Moscow, Russia

This study contributes to the research field of multimodal linguistics. Multimodal linguistics explores numerous channels involved in natural communication, such as verbal structure, prosody, gesticulation, mimics, eye gaze, etc., and treats them as parts of an integral process. Among the key issues in multimodal studies is the question of temporal coordination between the illustrative manual gestures (that is, spontaneous co-speech gestures) and elementary discourse units (that is, basic quanta of the local structure of spoken discourse). We address this issue with the help of a novel multimodal corpus "Pear chats and stories" that is currently under construction. It had been shown in a number of studies that gesture onset usually precedes speech onset. In order to verify this claim through our materials, we developed an analytic method that allowed to conduct a more detailed study. According to our results, it is only less than a half of all gestures that are produced before the corresponding fragment of talk. The most likely explanation of the obtained results is associated with gestures' affiliation in a certain functional class, that is strongly dependent on discourse genre and speakers' individual differences.

Key words: multimodal discourse, gesture, elementary discourse unit, temporal coordination

1. Мультимодальная коммуникация и мультимодальные корпуса

В доминирующей лингвистической традиции общепринято представление, согласно которому основным или даже единственным компонентом языкового взаимодействия является вербальная структура, а другие типы сигнала — в частности, просодия и жесты, — играют второстепенную роль и не относятся к собственно языку. В последние годы, однако, эта точка зрения постепенно

уступает место новой мультимодальной[2] перспективе (Scollon 2006; Кибрик 2010; Knight 2011; Abuczki, Esfandiari 2013; Müller et al. eds. 2014), согласно которой для успешной языковой коммуникации важны и, следовательно, заслуживают внимания все каналы передачи информации: вербальные единицы, просодия, жесты, мимика, направление взора и т. д. В нашем мультимодальном подходе выделяются два вокальных (слуховых) канала — вербальный и просодический, а также группа кинетических (зрительных) каналов (Кибрик 2010). Под вербальным каналом мы понимаем весь речевой материал, который в конечном счете сводится к последовательности фонем. К просодическому каналу относятся несегментные аспекты звука — интонация, дискурсивные акценты, громкость, тембр и т. д. (Кодзасов 2009; Кибрик, Подлесская ред. 2009). К кинетическим каналам (иногда именуемым языком тела) принадлежат мануальные жесты, направление взора, позы и т. д. (Крейдлин 2002; Kendon 2004; McNeill 2005; Николаева 2013).

Мультимодальный корпус — это «аннотированное собрание скоординированной информации из разных коммуникативных каналов, включая речь, направление взора, мануальные жесты и язык тела, которое обычно создается на материале записей человеческого поведения» (Foster, Oberlander 2007: 307–308). В отличие от моноканальных (письменных, основанных на вербальном канале) и мономодальных (речевых, основанных на вербальном и просодическом каналах, принадлежащих к слуховой модальности) корпусов, уже имеющих свою традицию, параметры, по которым можно классифицировать мультимодальные корпуса, еще только вырабатываются. Ниже мы перечислим четыре из них. Самый большой заявленный *объем* мультимодального корпуса — AMI Meeting Corpus, составляет 100 часов (Carletta 2006), однако бóльшая часть данных представлена в нем в виде неразмеченных видеофайлов. ***Характер общения*** собеседников удобно изображать в виде шкалы от контролируемых экспериментов на левом краю до ничем не ограниченного общения на правом. На самом левом краю находится Czech Audio–Visual Speech corpus (Železny et al. 2006), созданный для тестирования системы распознавания речи и включающий 25 часов записи 65 испытуемых, читающих вслух по 200 предложений. Правее расположен Fruit Carts Corpus (Aist et al. 2012), в котором записано 240 видеороликов продолжительностью 4–8 мин. каждый. Испытуемые выполняли стандартное задание — инструктор давал раскладчику инструкции по раскладыванию карточек с нарисованными на них фруктами. Еще правее находится D64 corpus, собранный для изучения социального общения (Campbell 2009), а также InSight Interaction corpus (Brône, Oben 2015), включающий 15 диалогов по 20 мин. каждый. На самом правом краю находятся корпуса, созданные в традиции анализа бытовых диалогов (Mondada 2014). Кроме объема и характера общения, выделяются также такие параметры, как ***количество собеседников*** (2 vs. 3+) и ***среда общения*** (специально созданные условия для

[2] Термин «мультимодальность» опирается на принятое в психологии и нейрофизиологии понимание модальности как принадлежности сигнала к определенной сенсорной системе человека.

проведения записей vs. неподготовленная среда). Три последних параметра важно оценивать с точки зрения естественности коммуникации. Наиболее естественные данные собираются в ходе бытового общения трех и более собеседников в неподготовленной среде. В описываемом ниже корпусе собраны записи общения четырех участников в специально созданных условиях совместного решения некоторой когнитивной задачи.

2. Корпус «Рассказы и разговоры о грушах»

Описываемый корпус является частью более обширного корпуса, создаваемого в рамках проекта РНФ «Язык как он есть: русский мультимодальный дискурс». Корпус включает 24 записи общей продолжительностью около 10 часов и объемом около 110 тыс. словоупотреблений. В качестве стимульного материала при сборе корпуса был использован известный шестиминутный «Фильм о грушах» У. Чейфа (Chafe ed. 1980). Была разработана новая методика сбора материала. В каждой записи принимали участие четыре человека с заранее распределенными ролями. Три коммуниканта — Рассказчик, Комментатор и Пересказчик — участвовали в основной части записи, а четвертый — Слушатель — присоединялся в конце. Сначала Рассказчик и Комментатор смотрели каждый на своем ноутбуке фильм и старались как можно лучше запомнить сюжет и всевозможные детали фильма. Затем Рассказчик излагал содержание фильма Пересказчику, который фильма не видел. На следующем этапе Комментатор дополнял рассказ Рассказчика подробностями, о которых тот не сообщил, при необходимости исправлял его, а Пересказчик уточнял у двух других участников необходимые для последующего пересказа детали; это был этап обсуждения. Наконец, Пересказчик пересказывал содержание фильма Слушателю, который непосредственно перед этим входил в помещение, — это был второй пересказ. После этого Слушатель письменно фиксировал на бумаге услышанный пересказ. Таким образом, основная задача каждого участника заключалась в том, чтобы максимально подробно и понятно донести до других коммуникантов полученную информацию.

При записи речи использовался рекордер ZOOM H6 Handy Recorder с параметрами 96 kHz / 24 bit; речь каждого из трех говорящих записывалась на петличный микрофон SONY ECM–88B; кроме того, отдельно велась общая стереозапись с микрофона рекордера. Три промышленные видеокамеры JAI GO–5000M–USB с частотой 100 к/с и разрешением 1392x1000 записывали крупным планом каждого говорящего в формате mjpeg, который выгодно отличается от других отсутствием межкадрового сжатия. Видеокамера GoPro Hero 4 Black Edition (50 к/с и 2700×1500) записывала общий план. Для регистрации движений глаз были использованы две пары очков–айтрекеров Tobii Glasses II Eye Tracker с частотой 50 Hz и разрешением видеокамеры 1920×1080. Один из айтрекеров был надет на Рассказчика, второй на Пересказчика. Насколько нам известно, подобные айтрекеры еще не использовались при исследовании мультимодального дискурса.

3. Жест и ЭДЕ: опережение, синхронизация или отставание?

3.1. Общие положения

Настоящая работа посвящена одному из ключевых вопросов мультимодальных исследований — ***временно́й координации*** между иллюстративными мануальными жестами и элементарными дискурсивными единицами (ЭДЕ, EDU)[3]. Жесты рук, сопровождающие речь, носят спонтанный характер и не имеют закрепленной формы и фиксированной связи между означаемым и означающим (Николаева и др. 2015). ЭДЕ является базовой единицей локальной структуры дискурса, выделяется на основании преимущественно просодических критериев и прототипически соответствует одной клаузе (Кибрик, Подлесская ред. 2009; Kibrik 2015).

Вопрос о временно́й координации «жест — ЭДЕ» восходит к более масштабному вопросу о том, насколько жесты и речь связаны между собой в когнитивной системе человека. МакНилл утверждает, что жесты и речь одновременно активируются в одном общем источнике, и, следовательно, должны быть синхронизированы как на фонологическом уровне, так и на уровне семантики (выражают один концепт) и прагматики (выполняют одну прагматическую функцию) (McNeill 1992); на эти положения опирается популярная модель Sketch (de Ruiter 2000). С другой стороны, в модели Interface (Kita, Ozyurek 2003) жесты и речь планируются в разных модулях и, соответственно, никаких гипотез о жесторечевой синхронизации в рамках данной модели не выдвигается. Проведенные исследования пока не дают однозначного ответа, какой подход больше соответствует действительности. Так, в работе Ozyurek et al. 2007 авторы, использовав метод вызванных потенциалов мозга, показали, что правильнее говорить не о временно́й, а только о семантической синхронизации, в то время как в работе Loehr 2012 автор говорит о наличии прагматической, структурной и временно́й синхронизации.

Несмотря на отсутствие согласия в вопросе, существует ли общий источник жестикуляции и речи, исследователи сходятся в том, что обычно ***время начала жеста опережает время начала речи***. Эта гипотеза была выдвинута МакНиллом (McNeill 1992), а в последние годы подтверждена на материале английского (Loehr 2012), французского (Ferré 2010) и польского (Karpiński et al. 2009) языков. Одно из возможных объяснений этого феномена состоит в том, что общий когнитивный источник, находящийся на досемантическом уровне, одновременно запускает активацию как абстрактных семантических репрезентаций, так и более конкретных моторных. Однако время, которое тратится на поиск моторных репрезентаций, обычно оказывается меньше, чем время, необходимое для поиска семантических репрезентаций. Это объяснение

---

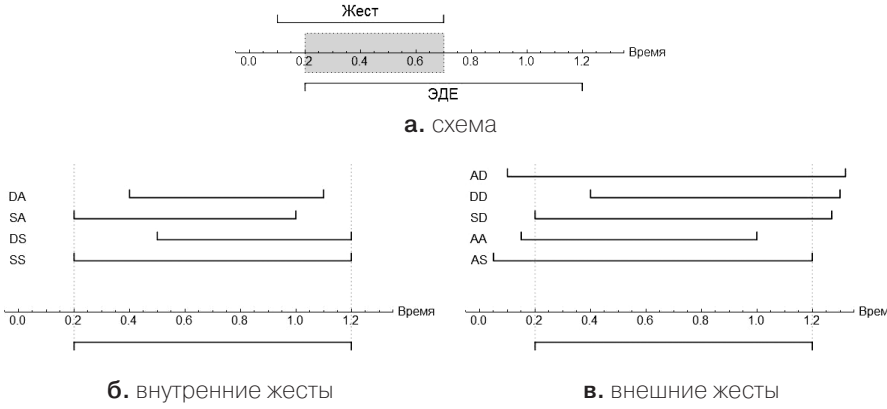
[3] При другом подходе к изучению временно́й координации исследуются отношения между ударными фазами жестов (stroke) и отдельными словами (Schegloff 1984; Leonard, Cummins 2009).

подтверждается в работе Morrel-Samuels, Krauss 1992: чем лучше известно говорящему некоторое слово, тем меньше временно́й интервал между началом жеста и соответствующего фрагмента речи.

Для проверки гипотезы об опережении жеста относительно ЭДЕ[4] мы использовали подкорпус, включающий четыре фрагмента по 6 мин. каждый (4 % всего корпуса); фрагменты были взяты из стадии обсуждения, так что в каждом была представлена речь всех трех собеседников. Аннотации речи и жестов были произведены независимо друг от друга. Для каждой из 1673 выделенных ЭДЕ в программе PRAAT (http://fon.hum.uva.nl/praat/) и для каждого из 614 жестов в программе ELAN (https://tla.mpi.nl/tools/tla-tools/elan/) с точностью до сотой доли секунды были определены время начала и время конца. Для определения временной координации была разработана специальная методика, реализованная в программе Wolfram Mathematica.

3.2. Термины и обозначения

Каждому жесту была поставлена в соответствие определенная ЭДЕ: жест считался соответствующим той ЭДЕ, с которой он имел наибольшее пересечение и которой соответствовал семантически. Длина жеста далее обозначена как L_g, длина ЭДЕ как L_{edu}; их общая часть, обозначенная заливкой на рис. 1, как $L_{g \cap edu}$.



а. схема

б. внутренние жесты

в. внешние жесты

Рис. 1. Временна́я координация «жест–ЭДЕ»

[4] При решении этой задачи необходимо определить, какая из двух единиц — жест или ЭДЕ — первична с точки зрения построения мультимодального дискурса. В работе Kibrik et al. 2015 показано, что ЭДЕ, определяемая на основе общих поведенческих критериев, является центральной единицей мультимодального дискурса. Таким образом, ниже мы рассматриваем вопрос о временно́й координации жеста *по отношению* к ЭДЕ.

Временно́й интервал между началом жеста и началом ЭДЕ мы назвали левым зазором (LGap), временно́й интервал между концом жеста и концом ЭДЕ — правым зазором (RGap). Если жест начинается не раньше ЭДЕ и заканчивается не позже, мы говорим о внутренних жестах (см. рис. 1а) и о внутренних левых (типы DS и DA[5]) и правых (SA и DA) зазорах. Если жест начинается раньше и/или заканчивается позже ЭДЕ, мы говорим о внешних жестах (рис. 1б) и о внешних левых (AS, AA и AD) и правых (SD, DD и AD) зазорах. Таким образом, мы выделяем девять логически возможных типов временно́й координации «жест–ЭДЕ».

При делении пар «жест — ЭДЕ» на типы важен вопрос о точности измерений. Что значит, что границы жеста совпадают с границами ЭДЕ (типы SS, DS, SA, AS и SD на рис. 1)? Мы производили данные вычисления с допустимой погрешностью (т. е. величиной зазора, признаваемого незначимым, далее Δ) в 200 мс, 100 мс и 50 мс, см. рис. 3 ниже.

Временна́я координация «жест — ЭДЕ» была количественно оценена при помощи следующих мер. Точностью (precision, P) мы называем отношение общей части к длине жеста: $L_{g \cap edu} / L_g$. Данная мера показывает, насколько точно жест вписывается в границы ЭДЕ. Полнота (recall, R) — это отношение общей части к длине ЭДЕ. Полнота показывает, насколько жест заполняет ЭДЕ: $L_{g \cap edu} / L_{edu}$. Среднее гармоническое (harmonic mean, HM) рассчитывается по формуле HM = 2PR / (P+R). Эту величину можно сравнить с F_1-мерой, которая используется в алгоритмах извлечения информации.

При исследовании вопроса о временно́й координации в целом мы используем все девять выделенных типов и все три меры, однако в данной работе для проверки гипотезы об опережении жеста относительно ЭДЕ в первую очередь мы учитываем деление жестов на внешние и внутренние, а также результаты измерения точности. Мы различаем ***относительную*** точность P, вычисляемую по вышеприведенной формуле, и ***абсолютную*** точность при измерении Δ в 200 мс, 100 мс и 50 мс. Очевидно, что при определении координации «жест–ЭДЕ» относительная и абсолютная точность могут не совпадать. Например, если $L_{g \cap edu}$ мало, то значение P также будет мало, но если при этом L_g меньше заданной Δ, то получается, что границы этого жеста и ЭДЕ совпадают. Следовательно, для большей надежности результатов нам необходимо использовать оба измерения.

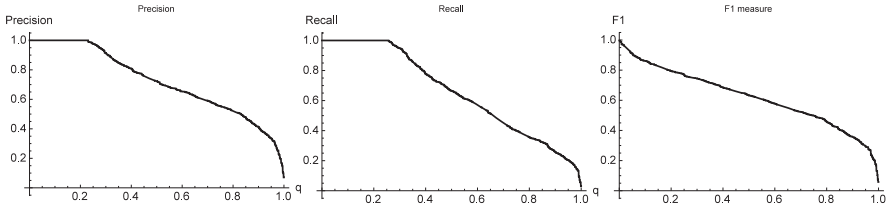
3.3. Результаты

Для каждой пары «жест — ЭДЕ» были вычислены L-Gap, R-Gap, границы $L_{g \cap edu}$, P, R и HM. Сводные данные по P, R и HM представлены на рис. 2.

По графикам видно, что ≈25 % жестов точно входят в границы ЭДЕ, а ≈85 % жестов входят в границы ЭДЕ с P ≥ 0.5, т. е. не менее половины жеста попадает внутрь ЭДЕ (рис. 2а); ≈25 % жестов полностью покрывают ЭДЕ, а 65 % жестов имеют R = 0.5, т. е. заполняют более половины ЭДЕ (рис. 2б); ≈20 % жестов имеют HM = 0.8, а 80 % жестов имеют HM = 0.5 (рис. 2в).

---

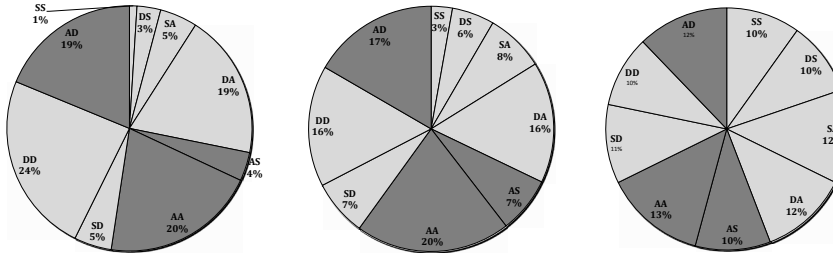
[5] от англ. Anticipation, Synchronization и Delay.



а. точность **б.** полнота **в.** гармоническое среднее

Рис. 2. Меры временно́й координации (по оси x отложены квантили, по оси y — значения соответствующей меры в диапазоне от 0 до 1)

Рассмотрим вопрос о временной координации с другой стороны, распределив пары «жест — ЭДЕ» по девяти типам, приведенным на рис. 1, с Δ в 200 мс, 100 мс и 50 мс, см. рис. 3.



а. Δ = 200 мс **б.** Δ = 100 мс **в.** Δ = 50 мс

Рис. 3. Распределение пар «жест — ЭДЕ» по типам

При Δ в 200 мс (рис. 3а) все типы представлены примерно в равных долях, а число внутренних жестов (SS, DS, SA и DA, 44 %) не намного меньше числа внешних. С уменьшением Δ до 100 мс (рис. 3б) и особенно до 50 мс (рис. 3в) распределение предсказуемым образом сдвигается в сторону преобладания тех типов, для которых не требуется строгое соответствие границ жеста и ЭДЕ (то есть типов с наличием одновременно LGap и RGap: DA, AA, DD и AD), а количество внутренних жестов уменьшается до 28 %, что уже сравнимо с P, равной 25 % (рис 1а). Таким образом, как с абсолютной, так и с относительной точностью только ≈¼ жестов вписывается в границы ЭДЕ. Можно предположить, что остальные ¾ жестов опережают ЭДЕ.

Чтобы проверить гипотезу об опережении жеста относительно ЭДЕ, выделенные типы были поделены на две группы:

(а) начало жеста опережает начало ЭДЕ (AS, AA и AD, на рис. 3 выделены темной заливкой);

(б) начало жеста синхронизировано или отстает от начала ЭДЕ (остальные типы).

Мы видим, что при Δ в 200 мс только 35% жестов попадает в группу (a); при Δ в 100 мс и 50 мс размер группы (a) увеличивается до 44% и 43%, соответственно, однако все равно не достигает 50%. Таким образом, полученные результаты не подтверждают гипотезу об опережающем производстве жестов.

4. Обсуждение результатов и перспективы дальнейших исследований

Итак, мы получили результаты, отличные от результатов предшественников. Какими причинами это может быть обусловлено? Ответить на этот вопрос можно, сравнив наш подкорпус с английским (Loehr 2012), французским (Ferré 2010) и польским (Karpiński et al. 2009) аналогами. Подчеркнем, что в каждой из этих работ авторы использовали просодические единицы, близкие к нашим ЭДЕ: Intonation Phrases в Ferré 2010, Intermediate Phrases в Loehr 2012 и Major Intonation Phrases (MajorIP) в Karpiński et al. 2009. Все эти понятия определяются на основе сравнимых просодических критериев и в общих чертах соответствуют друг другу.

Посмотрим сначала, насколько различаются результаты в целом. В нашем подкорпусе мы получили опережение жеста относительно ЭДЕ в 35% (200 мс), 44% (100 мс) и 43% (50 мс) случаев. Во французском корпусе опережают речь 70% жестов, в польском — 69%, для английского корпуса данные не приведены. Очевидно, что различия весьма велики.

Размер нашего корпуса (24 мин., 1673 ЭДЕ и 614 жестов) заметно превосходит аналоги: английский корпус включал 164 с записи, французский — 244 жеста, а польский — 223 жеста и 773 MajorIP. Серьезным преимуществом нашей работы является также возможность покадрово аннотировать видеозаписи, записанные с частотой 100 к/с. Из трех обсуждаемых корпусов английский был записан с частотой 30 к/с, французский — 24 к/с, для польского корпуса данные не приведены. Что касается измерения Δ, то в нашем подкорпусе данные были проанализированы с Δ в 50 мс, 100 мс и 200 мс. К сожалению, три другие работы не содержат эксплицитных указаний на процедуру подсчета Δ, однако в Karpiński et al. 2009 указано, что в 5% всех случаев начало жеста опережало начало речи меньше, чем на 100 мс, а в 40% случаев — меньше, чем на 200 мс. Кроме того, в Loehr 2012 упоминается, что в среднем время начало жеста опережало время начала речи на 100 мс. По-видимому, абсолютная точность во всех исследованиях была одинаковой или по крайней мере сопоставимой. Таким образом, с точки зрения размера корпуса и точности проводимых измерений наши результаты обладают более высокой степенью валидности.

Мы полагаем, что наиболее правдоподобное объяснение полученных различий связано с функциональными классами жестов. Хорошо известно, что иллюстративные жесты сильно различаются между собой по функциям, которые они выполняют в процессе коммуникации. Согласно популярной классификации МакНилла их можно разделить на: (1) указательные жесты, выполняющие референцию к объекту; (2) иконические жесты, изображающие конкретные

объекты или действия; (3) метафорические жесты, представляющие абстрактные понятия; (4) ритмические жесты, выделяющие фрагменты речи (McNeill 1992); ср. также несколько иную классификацию в работе Николаева 2013.

Французский корпус включал только иконические жесты, остальные корпуса включали все типы жестов. Известно, однако, что распределение жестов по функциональным классам в каждом конкретном случае сильно зависит от жанра дискурса и индивидуальных особенностей говорящих. Так, в частности, ритмические жесты в корпусе МакНилла составляли 44.4% всех жестов (McNeill 1992), в корпусе из работы Theune, Brandhorst 2010 — 32.1%, а в корпусе Николаевой (2013) — всего 15%. Из этих цифр следует, что ритмические жесты, обычно короткие и максимально синхронизированные с речью, могут оказывать сильное влияние на результаты временно́й координации в корпусе в целом. Кроме того, ритмические жесты корректнее анализировать, рассматривая временные отношения между ударными фазами жестов и отдельными словами, а не координацию «жест–ЭДЕ». Известно также, что в диалогической речи количество ритмических жестов увеличивается (Bavelas et al. 1992). Таким образом, на наши результаты могла повлиять разница в количестве ритмических жестов, скоординированных с ударными фазами жестов.

Как представляется, задача развития данного исследования в ближайшем будущем состоит в: (1) увеличении размера подкорпуса; (2) временно́й привязке каждого слова; (3) учете внутренних особенностей ЭДЕ (их синтаксической, коммуникативной и интонационной структуры); (4) разделении жестов на функциональные типы; (5) разделении жестов на подготовительную, ударную и ретракционную фазы (Kendon 1980); (6) анализе временно́й координации пар «жест — ЭДЕ» и «ударная фаза–слово» по разным функциональным типам жестов и типам ЭДЕ. Так, по-видимому, при анализе указательных и ритмических жестов правильнее рассматривать пары «ударная фаза–слово», а при анализе иконических и метафорических жестов — пары «жест — ЭДЕ».

Литература

1. *Abuczki A., Esfandiari B. G.* (2013), An overview of multimodal corpora, annotation tools and schemes, Argumentum, 9, pp. 86–98.
2. *Aist G., Campana E., Allen J., Swift M., Tanenhaus M. K.* (2012), Fruit Carts: A Domain and Corpus for Research in Dialogue Systems and Psycholinguistics, Computational Linguistics, 38 (3), pp. 469–478.
3. *Bavelas J. B., Chovil N., Lawrie D., Wade A.* (1992), Interactive gestures, Discourse Processes, 15 (4), pp. 469–489.
4. *Brône G., Oben B.* (2015), InSight Interaction. A multimodal and multifocal dialogue corpus, Language Resources and Evaluation, 49(1), pp. 195–214.
5. *Campbell N.* (2009), Tools and Resources for Visualising Conversational-Speech Interaction, in M. Kipp et al. (eds.) Multimodal Corpora: From Models of Natural Interaction to Systems and Applications, Springer, Heidelberg.

6. *Carletta J.* (2006), Announcing the AMI Meeting Corpus, The ELRA Newsletter, 11(1), January-March, pp. 3–5.
7. *Chafe W.* (ed.) (1980), The pear stories: Cognitive, cultural, and linguistic aspects of narrative production, Norwood, Ablex.
8. *Chafe W.* (1994), Discourse, consciousness, and time. The flow and displacement of conscious experience in speaking and writing, Chicago.
9. *de Ruiter J. P.* (2000), The production of gesture and speech, in D. McNeill (ed.), Language and Gesture. Cambridge University Press, pp. 248–311.
10. *Ferré G.* (2010). Timing Relationships between Speech and Co-Verbal Gestures in Spontaneous French, in Language Resources and Evaluation, Workshop on Multimodal Corpora, May 2010, Malta, pp. 86–91.
11. *Foster M. E., Oberlander J.* (2007), Corpus-based generation of head and eyebrow motion for an embodied conversational agent, Language Resources and Evaluation, 41 (3/4), pp. 305–323.
12. *Karpiński M., Jarmołowicz-Nowikow E., Malisz Z.* (2009), Aspects of gestural and prosodic structure of multimodal utterances in Polish task-oriented dialogues, Speech and Language Technology, 11, pp. 113–122.
13. *Kendon A.* (1980), Gesticulation and speech: Two aspects of the process of utterance, in M. R. Key (ed.), The relationship of verbal and nonverbal communication, pp. 207–227.
14. *Kendon A.* (2004), Gesture. Visible action as utterance. Cambridge.
15. *Kida T., Faraco M.* (2008), Prédication gestuelle, Faits de Langues, 31–32, pp. 217–226.
16. *Kibrik A. A., Podlesskaya V. I.* (eds.) (2009), Corpus of spoken Russian "Night Dream Stories" [Korpus ustnoy russkoy rechi "Rasskazy o snovideniyakh"], Jazyki slavyanskikh kul'tur, Moscow.
17. *Kibrik A. A.* (2010), Multimodal linguistics [Mul'timodal'naya lingvistika], in Yu. I. Aleksandrov, V. D. Solov'yev (eds.), Cognitive studies [Kognitivnyye issledovaniya], Vol. IV, Institute of psychology, Moscow, pp. 134–152.
18. *Kibrik, A. A.* (2015), The problem of non-discreteness and spoken discourse structure, Computational linguistics and intellectual technologies, 14 (21), vol. 1, pp. 225–233.
19. *Kibrik A., Fedorova O., Nikolaeva Ju.* (2015), Multimodal Discourse: In Search of Units, in G. Airenti, B. Bara, G. Sandini (eds.), Proceedings of the EuroAsian-Pacific Joint Conference on Cognitive Science, 4th European Conference on Cognitive Science, 11th International Conference on Cognitive Science, Torino, Italy, September 25–27, 2015, University of Torino, Torino, pp. 662–667.
20. *Kita S., Ozyurek A.* (2003), What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: evidence for an interface representation of spatial thinking and speaking, Journal of Memory and Language, 48 (1), pp. 16–32.
21. *Knight D.* (2011), Multimodality and active listenership: A corpus approach, Bloomsbury, London.
22. *Kreydlin G. E.* (2002), Nonverbal semiotics [Neverbal'naya semiotika], New literary review, Moscow.

23. *Leonard T., Cummins F.* (2009), Temporal alignment of gesture and speech, in E. Jarmołowicz-Nowikow, K. Juszczyk, Z. Malisz, M. Szczyszek (eds.), Proceedings of GESPIN2009: Gesture and Speech in Interaction, Poznan, Poland, pp. 1–6.
24. *Loehr D.* (2012)б Temporal, structural, and pragmatic synchrony between intonation and gesture, Laboratory Phonology, vol. 3 (1), pp. 71–89.
25. *McNeill D.* (1992), Hand and Mind: What Gestures Reveal about Thought, The University of Chicago Press, Chicago.
26. *McNeill D.* (2005), Gesture and thought, Chicago.
27. *Mondada L.* (2014), Bodies in action, Language and Dialogue, 4 (3), pp. 357–403.
28. *Morrel-Samuels P., Krauss R. M.* (1992), Word familiarity predicts temporal asynchrony of hand gestures and speech, Journal of Experimental Psychology: Human Learning and Memory, 18 (3), pp. 615–622.
29. *Müller C., Fricke E., Cienki A., McNeill D.* (eds.) (2014), Body — Language — Communication, Mouton de Gruyter, Berlin.
30. *Nikolaeva Yu. V.* (2013), Gesticulation in Russian discourse [Illustrativnyye zhesty v russkom diskurse]. Diss. cand. philol. science, MSU, Moscow.
31. *Nikolaeva Yu. V., Kibrik A. A., Fedorova O. V.* (2015), Discourse structure: a perspective from multimodal linguistics, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2015" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii], RGGU, Moscow, pp. 487–499.
32. *Ozyurek A., Willems R. M., Kita S., Hagoort P.* (2007), On-line integration of semantic information from speech and gesture: insights from event-related brain potentials, Journal of Cognitive Neuroscience, 19 (4), pp. 605–616.
33. *Schegloff E. A.* (1984), On some gestures' relation to talk, in J. M. Atkinson, J. Heritage (eds.), Structures of Social Action, Cambridge University Press, Cambridge, 266–298.
34. *Scollon R.* (2006), Multimodality and the language of politics, in K. Brown (ed.) Encyclopedia of language and linguistics, Elsevier, vol. 9, pp. 386–387.
35. *Theune M., Brandhorst C.* (2010), To beat or not to beat: beat gestures in direction giving, in S. Kopp, I. Wachsmuth (eds.), Gesture in Embodied Communication and Human–Computer Interaction, in Lecture Notes in Artificial Intelligence, vol. 5934. Springer, pp. 195–206.
36. *Železný M., Krňoul Z., Císař P., Matoušek J.* (2006), Design, implementation and evaluation of the Czech realistic audio-visual speech synthesis, Signal Processing, 83 (12), pp. 3657–3673.

STYLE AND GENRE CLASSIFICATION BY MEANS OF DEEP TEXTUAL PARSING

Galitsky B. A. (bgalitsky@hotmail.com),
Ilvovsky D. A. (dilvovsky@hse.ru),
Chernyak E. L. (echernyak@hse.ru),
Kuznetsov S. O. (skuznetsov@hse.ru)

National Research University Higher School of Economics, Moscow, Russia

In this paper we show that using deep textual parsing, which is finding complex features such as syntactic and discourse structures of the text, helps to improve the quality of style and genre classification. These results confirm achievements of many researches that have many times stated that using syntactic or morphological pattern for style and genre classification results in poor precision and recall. The best practice so far is to use n-gram patterns for this type of text classification problem. Syntactic and discourse structures allow however to capture some style of genre specific pattern of texts and to reach average precision higher than 95% on binary multi-genre classification.

Keywords: text genre, genre classification, rhetoric structure, discourse

КЛАССИФИКАЦИЯ ПО СТИЛЮ И ЖАНРУ С ИСПОЛЬЗОВАНИЕМ ДЕТАЛЬНОГО РАЗБОРА ТЕКСТА

Галицкий Б. А. (bgalitsky@hotmail.com),
Ильвовский Д. А. (dilvovsky@hse.ru),
Черняк Е. Л. (echernyak@hse.ru),
Кузнецов С. О. (skuznetsov@hse.ru)

Национальный исследовательский университет «Высшая школа экономики», Москва, Россия

Ключевые слова: жанр текста, стиль текста, риторические структуры, дискурс, мета-язык

1. Introduction

The problem of genre classification (also referred as automatic genre identification, AGI) has received so far some attention of the researches. Mainly there are two tied directions of these studies:

1. To develop intelligible genre system and to collect a corpus which would represent the established genre system. Usually the texts are collected from the Web [8, 11].
2. To develop a machine classifier for classifying texts of different genres [9–14].

In this paper we will consider both style and genre classification, without paying a lot of attention on the difference between these notions. Following [1] we will refer to "style" as to specific usage of language, and to genre as to the category of the text, which represent its intention and aim.

It is usually said that there are several applications of genre classification:

Evaluating how many different texts are there on the Web. This application can be treated as developing a socio- or psycho-metric tool [8, 11, 12, 13].

Using genre classification for improving user-based information retrieval: based on the query the search system should provide documents of appropriate genre (for example, if the query sounds scientific enough, return scholar papers, if the query is less formal—blogs, social media) [9].

Besides there are different attempts to genre classifications the majority of researches agree upon the following idea: the less complicated text elements are used as the features for classification, the better the results are. For example, [14, 28] suggest using character n-grams to perform genre classification on Brown corpus, BNC, HGC and other corpora. In [12] the syntactic patterns, morphological patterns and character n-grams are used to build feature sets and are compared to each other. The latter allow us to achieve the highest F-measure, while the former provide with poor results. The morphological pattern based classifier does not outperform the character-based one. In [13] common words are used to form feature sets.

To perform text classification in the described domains, we employ discourse information such as anaphora, rhetoric structure, entity synonymy. Relying on syntactic parse trees would provide us with specific expressions and phrasings connected with a style of writing. However, it will still be insufficient for a thorough description of linguistic features inherent to a style of writing. It is hard to identify such features without employing a discourse structure of a document. This discourse structure needs to include anaphora and rhetoric relations. Furthermore, to systematically learn these discourse features associated with the style of writing one needs a unified approach to classify graph structures at the level of paragraphs [16].

The design of such features for automated learning of syntactic and discourse structures for classification is still done manually today. To overcome this problem, tree kernel approach has been proposed [27]. Tree kernels constructed over syntactic parse trees, as well as discourse trees [17] is one of the solutions to conduct feature engineering. Convolution tree kernel [25] defines a feature space consisting of all subtree types of parse trees and counts the number of common subtrees to express the respective distance in the feature space.

The kernel ability to generate large feature sets is useful to assure we have enough linguistic features to differentiate between the classes, to quickly model new and not well understood linguistic phenomena in learning machines. However, it is often possible to manually design features for linear kernels that produce high accuracy and fast computation time whereas the complexity of tree kernels may prevent their application in real scenarios. SVM [20] can work directly with kernels by replacing the dot product with a particular kernel function. This useful property of kernel methods, that implicitly calculates the dot product in a high-dimensional space over the original representations of objects such as sentences, has made kernel methods an effective solution to modeling structured linguistic objects [26].

In this paper we will try to show how using more complicated and extensive syntactical information allows to improve the result of genre classification. The goal of this research is to apply the learning based on high-level linguistic features for the style and genre classification task and also to estimate the influence of the corpus annotation quality to the quality of the performance.

2. Style and genre classification

Moving from "simple" to "complex" system of style classes we start to distinguish texts between 2 classes: description (object-level) and meta-description (meta-language or meta-level). We consider domain of technical documents. In technical domain the description can be defined as a document which contains a thorough and well-structured text of how to build a particular system or work of art, from engineering to natural sciences to creative art. One can read such document and being proficient in the knowledge domain, can build such a system or work of art.

Conversely, meta-descriptions are explaining how to write particular description documents. They include manuals, standard documents should adhere to, tutorials on how to improve them, and others.

For the genre classification we used the system of genres and the corpus from [2, 3]. Let us describe the genre system in more details. Unlike other researches authors there do not define particular genres in crisp way, but constructs 17 main dimensions, so-called, Functional Dimension, in which a genre might be described. For example, the direction A7 corresponds to instructions (Tutorials, FAQs, manuals, recipes), the direction A11—to personal writing, such as diary-like blogs, personal letters, traditional diaries. A collection of texts, picked from the Web, is annotated by humans according to these directions: the annotator is asked to what extent this or that direction is present in the text. There are four possible answers: 0 none or hardly at all; 0.5 slightly; 1 somewhat or partly; 2 strongly or very much so. After the annotation, every text is represented as a vector in the space of 17 functional dimensions, which makes any kind of machine learning applicable. The texts and functional dimension are biclustered and the resulting clusters are said to represent a genre. The resulting system of genres consists of combinations of FTDs. Let us describe some of genres, achieved in [2,3]. There are genres that use only singly dimension: for example, the cluster Cl6 corresponds to the dimension A16, which is aimed at presenting

information. But the are some genres that correspond to two or three dimensions: the cluster Cl13 stands for dimensions A1 + A11, which are opinion blogs, often reporting personal experience and expressing one's emotions (43); and the cluster Cl14 stands for dimensions A11 + A19 + A3, which are diary blogs expressing one's emotions and attempting to embellish the description. The clusters often correspond to traditional genres, but are more reliable than traditional genres, since the annotator does not have to choose between several predefined genres. We adopt both the genre system and the corpus from this research.

3. Discourse text structure for the classification task

It turns out that low-level features could be insufficient for the genre classification in some domains like meta-document or design-document text detection. Since important phrases can be distributed through different sentences, one needs a sentence boundary—independent way of extracting both syntactic and discourse features. Therefore we intend to combine/merge parse trees to make sure we cover all the phrase of interest.

Rhetorical Structure Theory (RST) [5, 21] has been used to describe or understand the structure of texts and to link rhetorical structure to other phenomena, such as anaphora or cohesion. RST is one of the most popular approach to model extra-sentence as well as intra-sentence discourse. RST represents texts by labeled hierarchical structures. Their leaves correspond to contiguous Elementary Discourse Units; adjacent ones are connected by rhetorical relations (e.g., Elaboration, Contrast), forming larger discourse units (represented by internal nodes), which in turn are also subject to this relation linking. Discourse units linked by a rhetorical relation are further distinguished based on their relative importance in the text: nucleus being the central part, whereas satellite being the peripheral one. Discourse analysis in RST involves two subtasks: discourse segmentation is the task of identifying the EDUs, and discourse parsing is the task of linking the discourse units into a labeled tree.

Let us analyze how rhetoric relations could be useful in discriminating the writing style in the design-document domain. Let us consider the following piece of text.

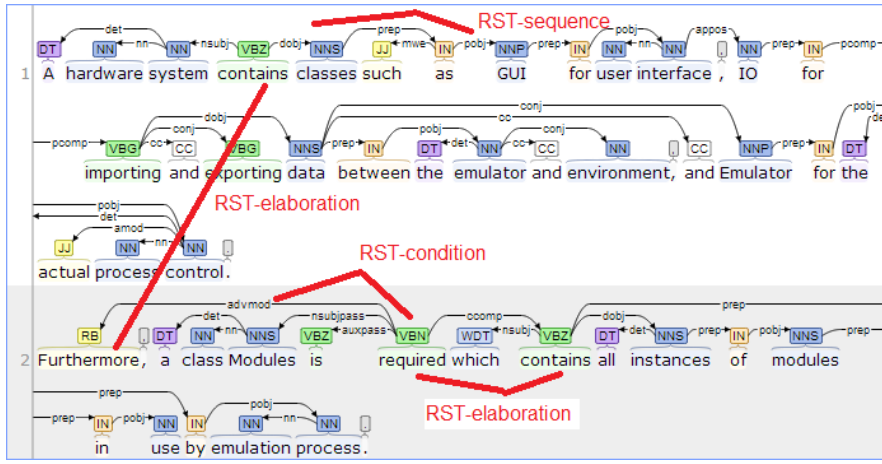
> *This document describes the design of back end processor. Its requirements are enumerated below.*

From the first sentence, it looks like an action-plan document. To process the second sentence, we need to disambiguate the preposition 'its'. As a result, we conclude from the second sentence that it is a requirements document, not an object-level action-plan one.

Discourse analysis explores how meanings can be built up in a communicative process, which varies between a text metalanguage and a text language-object. Each part of a text has a specific role in conveying the overall message of a given text.

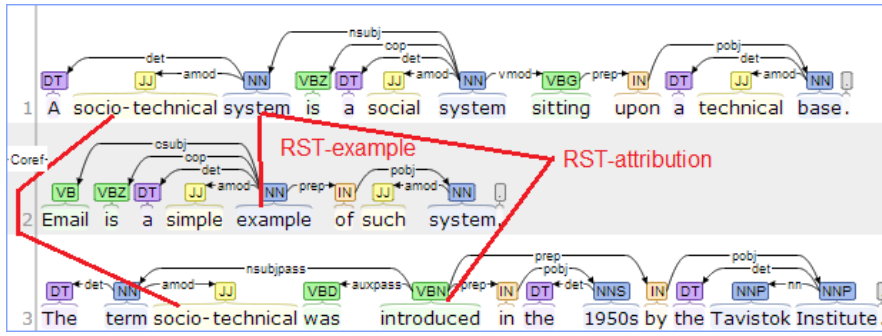
For the style classification tasks, just an analysis of a text structure can suffice for proper classification. Given a sequence from the [design-doc] class

A hardware system contains classes such as GUI for user interface, IO for importing and exporting data between the emulator and environment, and Emulator for the actual process control. Furthermore, a class Modules is required which contains all in-stances of modules in use by emulation process.



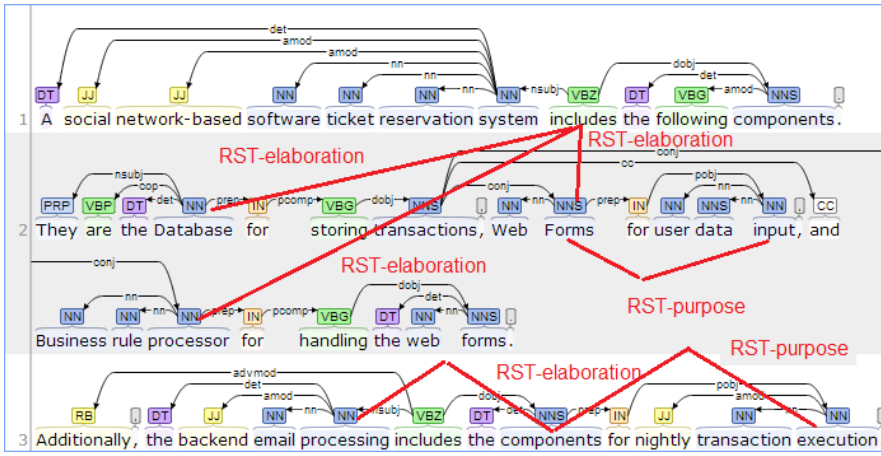
and a sequence from the [instruction] class

A socio-technical system is a social system sitting upon a technical base. Email is a simple example of such system. The term socio-technical was introduced in the 1950s by the Tavistok Institute.



We want to classify the following paragraph:

A social network-based software ticket reservation system includes the following components. They are the Database for storing transactions, Web Forms for user data input, and Business rule processor for handling the web forms. Additionally, the backend email processing includes the components for nightly transaction execution.



One can see that it follows the rhetoric structure of the top training set element, although it shares more common keywords with the bottom element. Hence we classify it as an design-doc, being an object-level text, since it describes the system rather than introduces a terms (as the bottom element does).

4. Learning on extended parse trees

The design of discourse and syntactic features for automated text assessment tasks is still an art nowadays. One of the solutions to systematically treat these features is the set of tree kernels built over syntactic parse trees, extended by discourse relations. Convolution tree kernel [25] defines a feature space consisting of all subtree types of parse trees and counts the number of common subtrees as the syntactic similarity between two parse trees. They have found a number of applications in a number of NLP tasks.

To obtain the inter-sentence links, we employed anaphoric relations from Stanford NLP [23, 24]. Rhetoric parser [16] builds a discourse parse tree by applying an optimal parsing algorithm to probabilities obtained from two conditional random fields, intra-sentence and multi-sentence parsing. We also rely on additional tags to extend SVM feature space, finding similarities between trees. These additional tags include noun entities from Stanford NLP such as organization and title, and verb types from VerbNet.

For every arc which connects two parse trees, we obtain the extension of these trees, extending branches according to the arc. For a given parse tree, we will obtain a set of its extension, so the elements of kernel will be computed for many extensions, instead of just a single tree [18]. The problem here is that we need to find common sub-trees for a much higher number of trees than the number of sentences in text, however by subsumption (sub-tree relation) the number of common sub-trees will be substantially reduced. The resultant trees are not the proper parse trees for a sentence, but nevertheless form an adequate feature space for tree kernel learning.

5. Evaluation

5.1. Style datasets

For the technical document domain, we formed a set of 940 description documents from the web. We also compiled the set of meta-documents on similar engineering topics, mostly containing the same keywords. We split the data into 3 subsets for training/evaluation portions and cross-validation.

Table 1. Evaluation results for technical documents

Method	Precision	Recall	F-measure
Nearest neighbor classifier (TF*IDF based)	53.9	62	57.67
Tree kernel—regular parse trees	71.4	76.9	74.05
Tree kernel SVM—extended trees for both anaphora and RST	**83.3**	**83.6**	**83.45**

Table 1 shows evaluation results. Baseline approaches show rather low performance. The one of the tree kernel based methods improves as the sources of linguistic properties are expanded. For both domains, there is an improvement by a few percent due to the rhetoric relations compared with the baseline tree kernel SVM which employs parse trees only. But for both domains the best accuracy is lower than 85%. This can be explained by a few reasons. Meta-documents can contain object-level text, such as design examples. Object level documents (genuine action-plan docs) can contain some author reflections on the system design process (which are written in metalanguage). Hence the boundary between classes does not strictly separates metalanguage and language object. So for better performance we need better annotated dataset.

5.2. Genre dataset

As it was mentioned earlier we adopted the genre system and the corpora from [1, 3]. The genre system is constructed in the following way. First, the Functional Text Dimensions (FTD) are defined. The FTD are genre annotations which reflect judgements as to what extent a text can be interpreted as belonging to a generalized functional category. A genre is a combination of several FTD. In other words, the genre is a point in the space, defined by FTD.

The corpus was annotated by humans. Every user was asked to evaluate texts of FTD on a scale: 0 none or hardly at all; 0.5 slightly; 1 somewhat or partly; 2 strongly or very much so. See an example of FTD and annotated texts below.

Table 2. Functional Text Dimensions

A1	Argum	To what extent does the text argue to persuade the reader to support (or renounce) an opinion or a point of view? ('Strongly', for argumentative blogs, editorials or opinion pieces)
A4	Fictive	To what extent is the text's content fictional? ('None' if you judge it to be factual/informative.)
A7	Instruct	To what extent does the text aim at teaching the reader how something works? (For example, a tutorial or an FAQ)
A8	Hardnews	To what extent does the text appear to be an informative report of events recent at the time of writing?
A9	Legal	To what extent does the text lay down a contract or specify a set of regulations?

In [2,3] twenty general dimensions are defined. Among them ten A1, A3, A4, A5, A6, A7, A8, A9, A11 form 7 different genres are defined. See the explanations of these genres bellow. For further classification we will exploit these genres.

- [tells] Instructions for how to use software.
- [tele] Instructions for how to use hardware.
- [ted] Emotional speech on a political topic. Presentation of him/her self. Attempt to sound convincing.
- [synd] An article on a political event by a professional journalist.
- [news] A presentation of a news article in an objective, independent manner.
- [fict] Novels, stories, verses.
- [un] UN reports.

Table 3. Main genres used for the evaluation

Genre example	A1	A3	A4	A5	A6	A7	A8	A9	A11
ted/eva_zeisel_on_the_playful_search_for_beauty	1	0	0	1	0	0	0	0	2
FictDostoyevskyF_CrimePun_II2_EN.txt	0	1	2	1	0.5	0	0	0	1
NewsGoalcom_MessiTop50_EN.txt	0.5	1	0	1	1	0	2	0	0.5
syndicate/exchange-rate-disorder	2	0	0	0	0	0.5	0.5	0	0
un/A_AC252_L13	1	0	0	0	0	0.5	0	2	0
TeleHTC_Manual_12_EN.txt	0	0	0	0	0	2	0	0	0
TelsGoog_Answer_2feb_EN.txt	0	0	0	0	0	0.5	0	0	1

Table 4. Pairwise classification results

Classes	VCDim	Recall	Precision	#kernel evaluations	F
Fict vs News	106	98.11	95.55	159,841	96.81
Ted vs Synd	787	99.49	98.94	73,177,349	99.21
Un vs News	697	98.70	94.93	9,486,134	96.78
Tele vs Tells	360	96.69	90.76	1,151,517	93.63
Fict vs Ted	139	97.12	93.74	7,557,291	95.40
Fict vs Synd	192	95.21	94.23	7,546,911	94.72
Fict vs Un	214	94.90	95.71	4,641,983	95.30
Fict vs Tele	317	97.25	94.90	6,547,910	96.06
Fict vs Tells	301	96.51	95.61	8,766,391	96.06
News vs Ted	514	96.85	93.85	2,619,549	95.33
News vs Synd	281	97.28	96.19	7,490,174	96.73
News vs Tele	190	96.31	94.27	5,235,193	95.28
News vs Tells	231	98.28	96.15	3,916,727	97.20
Ted vs Un	390	96.45	97.03	5,836,394	96.74
Ted vs Tele	210	97.28	96.62	1,612,102	96.95
Ted vs Tells	187	94.52	96.06	7,645,104	96.81

The values of quality measures—recall, precision and F-measure—are optimistically high. The highest F-measure is achieved by classification of Ted against Synd. Both of these genres correspond to describing political topics. However the rhetorical structures for these genres are completely different. Hence we are able to learn a very efficient classifier to distinguish between these genres.

Another important point is very impressive performance in the comparison with the results for the shallow-annotated dataset. Although the classes from this dataset could be roughly mapped on some genres (e.g. meta-level literature texts are corresponding with the [fict] genre) the distinction is less accurate.

6. Conclusions

We observed that using SVM TK one can differentiate between a broad range of text styles and genre. Each text style and genre has its inherent rhetoric structure which is leveraged and automatically learned. Since the correlation between text style and text vocabulary is rather low, traditional classification approaches which only take into account keyword statistics information could lack the accuracy in the complex cases.

In this paper we have presented three experiments on style and genre classifications. For the genre classification task we adopted a corpus annotated with 7 different genres and conducted a series of pairwise classification between two genres. From mathematical point of view, as a part of future extension of this research we may conduct one genre against all-others-genres-together classification, which will allow us to understand how distinctive each genre is. Hence we will obtain a more complete

picture of the genre system in general. If every genre is distinctive enough, it means that the whole genre system is well developed and the dimensions are adequate. However there might arise some problems because of the corpus being unbalanced: there are different numbers of texts if every genre and to tackle this problem we will have to balance the corpus.

References

1. *Lee, David YW.* Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. (2001)
2. *Sharoff, S.* In the garden and in the jungle: Comparing genres in the BNC and Internet. In Genres on the Web: Computational Models and Empirical Studies, pages 149–166. Springer, Berlin/New York. (2010)
3. *Sharoff, S., Wu, Z., and Markert, K.* The Web library of Babel: evaluating genre collections. In Proc. of the Seventh Language Resources and Evaluation Conference, LREC 2010, Malta. (2010)
4. *Richard Forsyth and Serge Sharoff.* Document Dissimilarity within and across Languages: a Benchmarking Study. Literary and Linguistic Computing, 29:6–22. (2014)
5. *Mann, W., Matthiessen, C., Thompson, S.:* Rhetorical Structure Theory and Text Analysis. Discourse Description: Diverse linguistic analyses of a fund-raising text / ed. by W. C. Mann and S. A. Thompson.—Amsterdam.—P. 39–78 (1992)
6. *Egg, M., Redeker, G.:* Underspecified discourse representation. In: Anton Benz & Peter Kühnlein (eds), Constraints in Discourse (pp. 117–138), Amsterdam: Benjamins. (2008)
7. *Taboada, M.:* The Genre Structure of Bulletin Board Messages. Text Technology 13 (2): 55–82. (2004)
8. *Biber, Douglas, Jerry Kurjian.* Towards a taxonomy of web registers and text types: a multidimensional analysis. Language and Computers 59.1 (2006): 109–131. (2006)
9. *Freund, Luanne, Charles LA Clarke, Elaine G. Toms.* Towards genre classification for IR in the workplace. Proceedings of the 1st international conference on Information interaction in context. ACM (2006)
10. *Kessler, Brett, Geoffrey Numberg, Hinrich Schütze.* Automatic detection of text genre. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics (1997)
11. *Santini, Marina.* Automatic identification of genre in web pages. Diss. University of Brighton (2007)
12. *Sarawgi, Ruchita, Kailash Gajulapalli, Yejin Choi.* Gender attribution: tracing stylometric evidence beyond topic and genre. Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics (2011)

13. *Stamatatos, Efstathios, Nikos Fakotakis, George Kokkinakis.* Text genre detection using common word frequencies." Proceedings of the 18th conference on Computational linguistics-Volume 2. Association for Computational Linguistics (2000).
14. *Wu, Zhili, Katja Markert, and Serge Sharoff.* Fine-grained genre classification using structural learning algorithms. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics (2010).
15. *Joty, S., Carenini, G., Ng, R., Mehdad, Y.:* Combining Intra- and Multi-sentential Rhetorical Parsing for Document-level Discourse Analysis.In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), Sofia, Bulgaria (2013)
16. *Galitsky, B., Ilvovsky, D., Kuznetsov, S. O., Strok, F.:* Matching sets of parse trees for answering multi-sentence questions // Proceedings of the Recent Advances in Natural Language Processing, RANLP 2013.—INCOMA Ltd., Shoumen, Bulgaria.—P. 285–294 (2013)
17. *Ilvovsky, D.:* Going beyond sentences when applying tree kernels. Proceedings of the Student Research Workshop.—ACL 2014.—P. 56–63 (2014)
18. *Galitsky, B., Kuznetsov, S. O.:* Learning communicative actions of conflicting human agents. J. Exp. Theor. Artif. Intell. –Vol. 20(4).—P. 277–317 (2008)
19. *Vapnik, V.:* The Nature of Statistical Learning Theory.—Springer-Verlag (1995)
20. *Marcu, D.:* From Discourse Structures to Text Summaries. Proceedings of ACL Workshop on Intelligent Scalable Text Summarization / eds. I. Mani and M. Maybury.—Madrid, P. 82–88 (1997)
21. *Severyn, A., Moschitti, A.:* Fast Support Vector Machines for Convolution Tree Kernels. Data Mining Knowledge Discovery 25.—2012.—P. 325–357. (1997)
22. *Marta Recasens, Marie-Catherine de Marneffe, and Christopher Potts.* The Life and Death of Discourse Entities: Identifying Singleton Mentions. In Proceedings of NAACL (2013)
23. *Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., Jurafsky, D.:* Deterministic coreference resolution based on entity-centric, precision-ranked rules. Computational Linguistics 39(4), (2013)
24. *Collins, M., Duffy, N.:* Convolution kernels for natural language. In Proceedings of NIPS, 625–632 (2002)
25. *Moschitti, A.:* Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees. In Proceedings of the 17th European Conference on Machine Learning, Berlin, Germany (2006)
26. *Sharoff, Serge.* Functional Text Dimensions for annotation of Web corpora. http://corpus.leeds.ac.uk/serge/publications/2015-corpora-submission.pdf (2015)
27. *Cumby, C. and Roth D.* On Kernel Methods for Relational Learning. ICML, pp. 107–14. (2003)
28. *Kanaris, I. and E.*Stamatatos. Learning to Recognize Webpage Genres Information Processing and Management, 45(5), pp. 499–512, Elsevier (2009)

ВИД РУССКОГО ГЛАГОЛА: ЖЕСТИКУЛЯЦИОННЫЙ ПРОФИЛЬ

Гришина Е. А. (rudi2007@yandex.ru)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

В статье на материале Мультимедийного русского корпуса (МУРКО) рассматриваются жестикуляционные параметры, которые позволяют различить СВ- и НСВ-глаголы. В качестве таковых выделены параметр длительности жеста, его кратности и его энергетики. Показано, что параметры кратности и энергетики при различения СВ- и НСВ-глаголов имеют производный характер (т. е. они различают СВ- и НСВ глаголы только потому, что хорошо проявляют себя внутри одной из групп и никак не проявляются себя в другой группе). В качестве же единственного параметра, который работает для различения СВ- и НСВ-глаголов, в русской жестикуляции выступает параметр длительности жеста.

Ключевые слова: вид русского глагола, жестикуляция, жест, устная коммуникация, Мультимедийный русский корпус (МУРКО)

RUSSIAN VERBAL ASPECT AND GESTICULATION

Grishina E. A. (rudi2007@yandex.ru)

Vinogradov Institute of Russian Language RAS, Moscow, Russia

In this paper, we use evidence from the Multimodal Russian Corpus (MURCO) to explore gesture properties that enable distinction between perfective and imperfective Russian verbs. The properties identified are duration, repetition, and energy. We show that repetition and energy differentiate perfective and imperfective verbs because these properties are salient in gestures accompanying one group of verbs but are not manifest in gestures accompanying verbs in the other group. Gesture duration, on the other hand, can be used to identify either aspect.

Key words: Russian verbal aspect, gesticulation, gesture, spoken interaction, Multimodal Russian Corpus (MURCO)

1. Введение

В своей лекции на «Диалоге'2015», прочитанной в качестве приглашенного докладчика, Алан Ченки посетовал на то, что в рамках руководимого им проекта РНФ «Взаимодействие вербальных и невербальных средств конструирования событий в разных языках» не удалось обнаружить каких-то специфических жестикуляционных событий, отражающих видовые противопоставления в русском языке. На тот момент (май 2015 г.) автор настоящей статьи в рамках уже своего исследования имел аналогичные — скорее отрицательные, чем положительные — результаты. Эта статья представляет собой попытку поставить некоторые новые «ловушки» на русский вид и попытаться предложить жестикуляционные параметры, которые позволят утверждать, что видовые противопоставления все-таки отражаются в русской жестикуляции.

Замечание. Основанием для оптимистических ожиданий послужила модель У. Крофта [Croft 2012], творчески развитая в работе [Янда 2015]: аспектуальные контуры здесь изображаются разными типами геометрических фигур, сочетающих в себе линии (часть ситуации, находящаяся в фокусе внимания) и точки и расположенных в системе координат, включающих в себя параметры времени и качественного состояния. С нашей точки зрения, лингвистическое явление, которое можно изобразить с помощью геометрической фигуры, не может не найти своего отражения в жестикуляции, хотя бы и в очень общем и трансформированном виде.

1.1. Методика работы

Из Мультимедийного русского корпуса (МУРКО) была отобрана база данных, включающая в себя в общей сложности порядка полутора тысяч жестоупотреблений, сопровождающих глагольные формы (личные и инфинитив; причастия в рассмотрение не включались)[1]. База данных включает в себя только бесприставочные глаголы, чтобы избежать влияния на результаты жестикуляционных профилей приставок (см. [Гришина 2013)]: глаголы СВ с суффиксом *ану-* (36 жестоупотреблений[2]), с суффиксом *ну* (443 жестоупотребления[3]), бессуфиксальные, имеющие только тематическую гласную, типа *бросить, решить* и под. (81), глаголы НСВ (в общей сложности 875 жестоупотреблений).

---

[1] О том, что означает в данном контексте глагол «сопровождать», будет показано несколько позже.

[2] К сожалению, этот замечательный класс глаголов относительно редко встречается в МУРКО ввиду довольно низкой доли сниженного регистра речи в корпусе. Но имеющихся примеров оказалось достаточно, чтобы выдвигать некоторые предположения.

[3] Далее мы будем позволять себе использовать — пусть и не вполне корректно — термин *семельфактивы* по отношению к СВ-глаголам с суффиксами *-ану-* и *-ну-* в совокупности.

Жесты, сопровождающие глагол, были описаны по ряду параметров. Часть параметров, как и ожидалось, оказалась незначимой (что и давало нам прежде основания утверждать, что русский вид в жестикуляции не отражается[4]), однако обнаружились факторы, которым мы ранее при описании русской жестикуляции не уделяли должного внимания, и вот они показали связь с русским видом, иногда довольно сильную.

1.2. Структура статьи

После описания используемых параметров будут проведены общие противопоставления СВ и НСВ, как обычно, с привлечением статистических данных. Кроме того, будут проведены некоторые разделительные линии уже внутри групп СВ- и НСВ-глаголов соответственно. Статистическим данным будут даны интерпретации.

2. Параметры, используемые для различения СВ vs. НСВ

2.1. Зона действия жеста и его длительность

Зоной действия мы называем ту часть фразы, которая сопровождается данным жестом, начиная с предударного удержания (если оно имеется), включая ударную часть жеста и заканчивая постударным удержанием (опять же, если таковое имеется)[5]. Фраза и ее части здесь (и далее) измеряются в фонетических словах[6]. Мы различаем зону жеста в одно фонетическое слово (см. рис. 1 и примеры (1)–(5)) и зону жеста, включающую более одного фонетического слова (см. рис. 2 и примеры (6)–(10)).

---

[4] Незначимые параметры (направление движения по трем декартовым координатам, типы движения головы, бóльшая часть конфигураций ладоней и траекторий, степень напряженности ладони) нами в статье описываться не будут.

[5] О фазах жеста см. [Kendon 1972, 1980], [Kita et al. 1998], [Bressem, Ladewig 2011].

[6] Измерение зоны действия жеста в фонетических словах, а не с привлечением объективных временны́х параметров типа секунд и миллисекунд избавляет нас от необходимости учитывать разную скорость говорения и жестикуляции как у разных говорящих, так и у одного и того же говорящего в разных ситуациях: приведение этих десемантизированных данных к общему знаменателю представляет собой отдельную проблему. Кроме того, поскольку жестикуляция очевидным образом ориентирована на смысл сопровождающей ее речи, обезличенные объективные временны́е показатели могут оказаться и вовсе бесполезными — соотношение жеста с фонетическими словами в этой ситуации выглядит гораздо более предпочтительным.

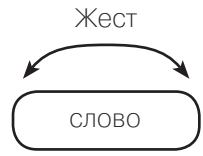


Рис. 1

(1) Одним словом, я рука вверх{рванул}... и вытянул себя из болота. munhauzen_005.mp4

(2) Я хочу рука вперед и вверх{швырнуть им} всем зажравшимся скотам nastrojsh_256.mp4

(3) Ты кивок{станешь} сам на время бездыханным korol-olen_095.wmv

(4) рука и голова направо{За ноги} хватать их? prost_veshi_028.mp4

(5) Ты стояла рука вниз{у окна}... otpusk_sentabre_163.wmv

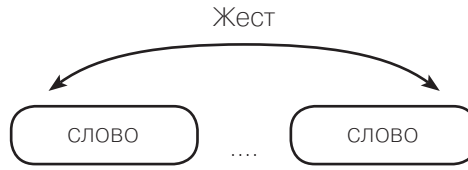


Рис. 2

(6) Так мы с полного хода кулак вперед налево с замахом и отскоком[7]{и резанули в причал носом} polosat_rejs_111.mp4

(7) рука слева направо и вниз{И вот он чиркнул...} gordon_aster_opasn_046.mp4

(8) рука к собеседнику{Простите, Глеб Максимильяныч} kommunist_063.wmv

(9) руки вверху изнутри наружу{можно сыпать именами} academ_basov_sredn_civiliz_022.mp4

(10) руки в конфигурации контейнер перед говорящим{сидит не шелохнется} vasil_prekras_010.wmv

Таким образом, выравнивание жеста с одним или с последовательностью фонетических слов характеризует длительность жеста, выраженную в фонетических словах. Далее мы будем называть жест длиной в одно фонетическое слово **кратким жестом**, в несколько фонетических слов — **удлиненным**.

---

[7] О *замахе* и *отскоке* см. ниже.

Замечание. Зона действия жеста может содержать 1) один глагол, 2) глагол и зависимые от глагола члены предложения, примыкающие к глаголу, 3) только зависимые от глагола элементы (см. примеры (4)–(5)). В третьем случае, следовательно, жест выровнен с зоной действия, не содержащей глагола. Как показал статистический анализ материала, в случае зоны действия, содержащей только зависимые члены, видовые различия никак в жестикуляции не проявляются, поэтому такой тип выравнивания мы просто вывели из рассмотрения. Это означает, что краткие жесты всегда выровнены только с глаголом, а удлиненные — с группой слов, одно из которых обязательно глаголом является.

2.2. Кратность жеста

Мы различаем однократные и многократные жесты. **Многократными** мы считаем:

1) Истинно многократные жесты, которые не могут быть в принципе осуществлены в однократном режиме, т. е. ударная часть которых не может содержать менее двух движений. Истинно многократными жестами являются, например, все типы колебательных движений (см. [Гришина 2015а]), поскольку они не могут осуществляться одним движением пальцев или руки, обязателен симметричный или асимметричный повтор движения. Истинно многократно также отрицательное качание головой.
2) Квантующие жесты (см. [Гришина 2015б]), т. е. жесты (обычно ручные, но изредка фиксируются и головные), прямолинейная траектория которых делится на равные отрезки (в частности, движением типа *циклоида*). Квантующим является также многократное круговое движение, которое можно рассматривать как циклоиду, не разложенную на кванты вдоль некоего вектора.
3) Жестикуляционный повтор, когда на одну экспозицию и ретракцию приходится более одной ударной части жеста.

Все остальные жесты считаются **однократными**.

Истинно многократные и квантующие жесты могут быть как краткими, так и удлиненными. Жестикуляционный повтор осуществляется на протяжении одного фонетического слова, а следовательно, может быть только кратким[8].

Иллюстрации: рис. 3, примеры (11)–(15) — однократные жесты, рис. 4, примеры (16)–(20) — многократные жесты.

---

[8] Жестикуляционные повторы, которые захватывают зону длительностью более одного фонетического слова, представляют собой либо жестикуляционное скандирование, либо фатическую многократность (см. [Гришина 2014]). Жесты этих типов не показывают никакой связи с видовыми противопоставлениями и поэтому в статье не рассматриваются.

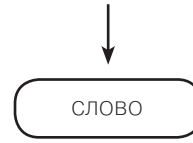


Рис. 3

(11) хотим <…> «Серапионовых братьев» кулак сверху вниз{долбануть} my_iz_djaza_147.mp4

(12) да как две руки сверху вниз{прыгнешь} в Черное две руки сверху вниз подхват{море}[9]! odn_letom_022.mp4

(13) Не придешь, кивок{брошу всё} и подхват{уеду} на производство. penkovo_019.wmv

(14) то рука вниз с отскоком{хватать их за руки} за ноги timur_komanda_109.wmv

(15) вам нужно рука вниз налево{больше} ходить liubovn_041.wmv

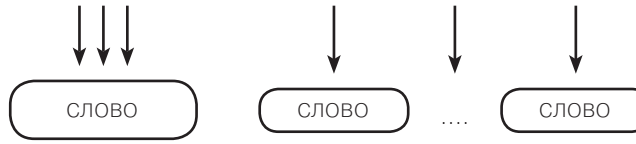


Рис. 4

(16) когда со Стенькой купца множественные движения головы из стороны в сторону{грабанули} на Волге sluga_098.mp4

(17) поэтому два электрона со сверхбольшими скоростями можно снаружи в центр дважды{стукнуть}… gordon_mir_vacuum_061.mp4

(18) оставила ей две штуки, попросила неоднократное круговое движение указательным пальцем{реализовать} interdevoch_375.wmv

(19) а хлопки в ладоши{ему все хлопают} zigzag_udachi_098.wmv

(20) вы должны многократные круговые движения открытой ладонью{менять свой облик} pokrovskie_vorota_265.wmv

---

[9] На словоформе *море* мы имеем дело с подхватом (catchment) предыдущего жеста на *прыгнешь* (о явлении подхвата см. [McNeill et al. 2001]).

2.3. Конфигурация ладони

Как показал анализ материала, существенным для различения видов оказалось противопоставление всего двух конфигураций ладони — *кулак* (см. примеры (6), (11)) и *открытая ладонь*.

2.4. Энергетика движения

Обращает на себя внимание параметр, который мы не привлекали к анализу ранее и который, как оказалось, имеет некоторое отношение к видовому противопоставлению. Условно-предварительно его можно назвать *энергетикой движения*. Выражается этот параметр двумя способами.

2.4.1. Замах

На стадии экспозиции ладонь занимает максимально периферийное положение, т. е. отводится говорящим на максимально отдаленное, но при этом все еще комфортное для жестикулирующего положение относительное центральной зоны жестикуляции. Тем самым рука на стадии экспозиции как бы заявляет максимальную **потенциальную энергию** (во вполне физическом значении этого термина), которую говорящий готов вложить в осуществляемый жест на ударной стадии (см. рис. 5).

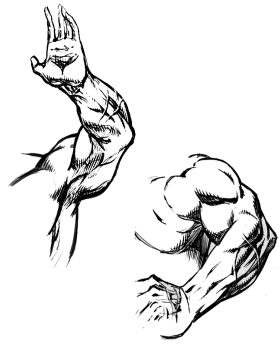


Рис. 5

Замах характерен в основном для движения рук, но встречается также при исполнении жестов головы, прежде всего, — вертикального кивка сверху вниз: кивок с замахом характеризуется максимальным отведением головы назад и вверх на стадии экспозиции, с последующим энергичным движением вниз на ударной стадии, ср. пример (21):

(21) Очень боюсь, что когда-нибудь они кивок{лопнут} pechki-lavoch_071.mp4

2.4.2. Отскок

После ударной стадии, на стадии ретракции ладонь возвращается несколько назад под некоторым углом к пройденной траектории (рука сгибается в локте), тем самым имитируя отскок некоторого предмета после сильного удара по поверхности (рис. 6).

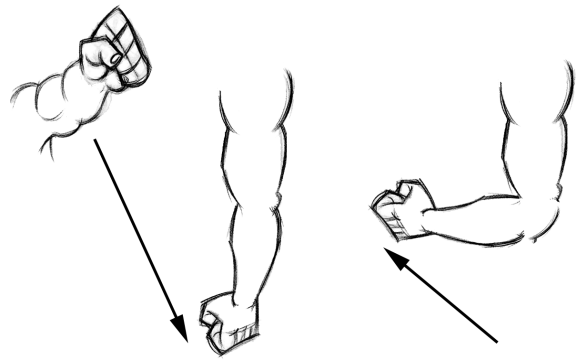


Рис. 6

Замах и отскок могут быть одновременно использованы при исполнении жеста, см. пример (6), а также:

(22) подкарауливать империалистов и [замах]{топить их} [подхват, замах и отскок]{как котят}, вот. drug_kolka_354_05.wmv

Замечание. Ранее, при обсуждении траекторий движения руки по поперечной оси, мы уже описывали такую траекторию, как радикальное пересечение, — движение на ударной стадии левой руки из крайней правой периферии налево или правой руки из крайней левой периферии направо. Эта траектория задает максимальное расстояние, которое рука может произвести на ударной стадии жеста. Радикальное пересечение характерно для жеста *ручное отрицание*, для изображения линий и поверхностей, но также оно с повышенной частотой сопровождает приставки, характеризующиеся повышенной энергетикой (*вы-, от-, у-*) (см. [Гришина 2013]). Но если радикальное пересечение передает метафору высокой энергетики опосредованно, через идею максимального расстояния, преодоленного жестикулирующей рукой, то такие характеристики жеста, как *кулак, замах* и *отскок*, передают идею энергии непосредственно.

3. Статистические данные

Приведем несколько таблиц, которые показывают влияние перечисленных выше факторов на противопоставление СВ и НСВ[10].

Таблица 1. Виды глагола и длительность жеста[11]

Длительность жеста / Вид	Краткий жест	Удлиненный жест
СВ	**441**	*84*
НСВ	*504*	**255**
$\chi^2 = 49{,}45$, $p = 2{,}03–12$, параметры связаны, распределения достоверны		

Таблица 2. Виды глагола и кратность жеста

Кратность жеста / Вид	Многократный	Однократный
СВ	*24*	476
НСВ	**73**	614
$\chi^2 = 13{,}09$, $p = 0{,}0003$, параметры связаны, распределения достоверны		

Таблица 3. Виды глагола и конфигурация ладони

Конфигурация ладони / Вид	*Кулак*	*Открытая ладонь*
СВ	**65**	*160*
НСВ	*61*	387
$\chi^2 = 22{,}96$, $p = 1{,}65–06$, параметры связаны, распределения достоверны		

Таблица 4. Виды глагола и энергетика движения

Тип движения / Вид	*Замах* и/или *отскок*	Стандартное движение
СВ	28	40
НСВ	532	835
$\chi^2 = 0{,}14$, $p = 0{,}71$, параметры не связаны, распределения недостоверны		

[10] В таблицах полужирным шрифтом и затемнением ячейки обозначены данные, которые существенно выше средних распределений, подчеркнутым курсивом — те данные, которые существенно меньше средних.

[11] В таблицу не включены данные по контекстам, включающим в себя жестикуляционное скандирование.

Итак, вы видим, что для различения **в общем** СВ и НСВ рабочими оказываются только три параметра — длительность, кратность и энергетика жеста, конфигурация ладони значения не имеет. При этом значения параметров распределены между видами так, как показано в табл. 5.

Таблица 5

Вид / Параметр	СВ	НСВ
Длительность	краткий	удлиненный
Кратность	не многократный	многократный
Энергетика	жест с повышенной энергетикой	стандартный жест

Для дальнейшего анализа следует произвести более детальное разделение внутри групп СВ- и НСВ-глаголов. В группе СВ-глаголов будем различать отдельно глаголы на *ну-*, глаголы на *ану-*, семельфактивы как группу глаголов с суффиксом (*а*)*ну-* и бесприставочные СВ-глаголы. В группе НСВ-глаголов будем различать мультипликативы (формальный признак — способность образовывать семельфактивы, *кидать — кинуть, чесать — чесануть*) и континуативы, или обозначения непрерывных процессов (формальный признак — неспособность к образованию глаголов с суффиксами (*а*)*ну-*, *бросать, думать, ходить*).

Проверим работу предложенных параметров на этих группах.

Таблица 6. Глаголы и длительность жеста

Вид / Длительность жеста	Краткий жест	Удлиненный жест
семельфактивы	**388**	*61*
НСВ	*504*	**255**
СВ бесприставочный	53	23
$\chi^2 = 49{,}45$, $p = 2{,}03–12$, параметры связаны, распределения достоверны		

Внутри группы *семельфактивы* данный параметр незначим, т. е. не различает глаголы на *ну-* и *ану-*. Аналогичным образом, этот параметр незначим внутри группы НСВ, т. е. не различает мультипликативы и континуативы (данные не приводим в целях экономии места).

Таблица 7. Глаголы и кратность жеста

Вид / Кратность жеста	Многократный жест	Однократный жест
глаголы с суффиксом *ану-*	3	31
глаголы с суффиксом *ну-*	20	374
СВ бесприставочный	1	71
$\chi^2 = 3{,}1$, $p = 0{,}21$, параметры не связаны, распределения недостоверны		

Кратность жеста / Вид	Многократный жест	Однократный жест
мультипликативы	**45**	*142*
континуативы	*28*	472
$\chi^2 = 48{,}9$, $p = 2{,}75–12$, параметры связаны, распределения достоверны		

Как видим, этот параметр незначим внутри группы СВ и при этом очень неплохо различает мультипликативы и континуативы в группе НСВ.

Таблица 8. Глаголы и конфигурация ладони

Конфигурация ладони / Вид	*Кулак*	*Открытая ладонь*
глаголы с суффиксом *ану-*	**10**	*7*
глаголы с суффиксом *ну-*	50	114
СВ бесприставочный	*5*	**39**
$\chi^2 = 14{,}2$, $p = 0{,}0008$, параметры связаны, распределения достоверны		
Конфигурация ладони / Вид	***Кулак***	***Открытая ладонь***
мультипликативы	17	90
континуативы	44	297
$\chi^2 = 0{,}62$, $p = 0{,}43$, параметры не связаны, распределения недостоверны		

Параметр *конфигурация ладони* различает глаголы с суффиксом *ану-* и бесприставочные СВ. Внутри группы НСВ данный параметр незначим.

Таблица 9. Глаголы и энергетика движения

Тип движения / Вид	*Замах и/или отскок*	Стандартное движение
глаголы с суффиксом *ану-*	**8**	28
глаголы с суффиксом *ну-*	18	425
СВ бесприставочный	2	79
$\chi^2 = 24{,}4$, $p = 5{,}05–06$, параметры связаны, распределения достоверны		
мультипликативы	9	209
континуативы	31	626
$\chi^2 = 0{,}13$, $p = 0{,}72$, параметры не связаны, распределения недостоверны		

Для группы НСВ данный параметр незначим, в группе СВ-глаголов он противопоставляет глаголы с суффиксом *ану-* всем остальным СВ-глаголам, для которых этот параметр незначим.

Подытожим сказанное в табл. 10.

Таблица 10

Параметр / Группа глаголов	Длительность жеста	Кратность жеста	Конфигурация ладони	Энергетика
глаголы с суффиксом *ану-*	краткий		*кулак*	*замах/ отскок*
глаголы с суффиксом *ну-*				
СВ бесприставочный			*открытая ладонь*	
мультипликативы	удлиненный	многократный		
континуативы		не многократный		

4. Возможные интерпретации

4.1. СВ vs. НСВ

Итак, первое, что мы видим из табл. 5, 10: в русской жестикуляции СВ противопоставлен НСВ как вид, обозначающий краткое (в пределе — нульмерное, мгновенное) событие или сверхкраткий процесс (*свертки*, по [Плунгян 2000]), — виду, обозначающему длительные, продолженные события или процессы. При это в зоне бесприставочных глаголов за краткость отвечают суффиксы (*а*)*ну-*: бесприставочные и бессуффиксальные СВ-глаголы, как мы видим из табл. 10, занимают промежуточное положение между семельфактивами и НСВ-глаголами: для них в жестикуляционном плане параметр длительности несуществен, и тем самым глаголы типа *бросить, решить, дать, стать* в жестикуляции позиционируют себя не как мгновенные события или очень короткие процессы, а как **завершенные** события/процессы[12]. Таким образом, в русской

[12] В связи с этим любопытно различие между глаголами *кинуть* и *бросить* (*бросить мяч*, но ?*кинуть мяч*): *кинуть* акцентирует начальный этап броска и пренебрегает его целью, *бросить* акцентирует не только начальную, но и конечную точку действия, а следовательно, для *бросить* становится актуальной и цель броска. Таким образом, в отличие от [László 2008], мы не считаем, что «глагол *бросить* (*мяч, камень*) в плане моментальности не отличается от синонимичных глаголов *кинуть* и *швырнуть*, в которых имеется суффикс *-ну-*»: для *кинуть, швырнуть* важно указание на моментальный характер действия, вне зависимости от того, завершилось ли это действие или осталось незавершенным, в то время как для *бросить* важно, что моментальное событие броска привело к завершению этого процесса (что в данном конкретном случае осмысливается как ответ на вопрос — была ли цель у данного броска, достижение которой можно было бы считать актом завершения данного действия, или цели не было, вернее, цель была только в том, чтобы кинуть или швырнуть, а не в том, чтобы мяч достиг задуманной конечной точки движения).

жестикуляции противопоставление СВ- и НСВ-глаголов хорошо иллюстрирует мысль, высказанную в [Плунгян 2000: 214]: «в русском языке видовая оппозиция вообще в значительно большей степени ориентирована на *длительность* ситуаций (длящиеся vs. мгновенные или краткие <…>), чем на ограниченность ситуаций временными рамками (как это свойственно, например, романским и германским языкам)». С этой точки зрения, конечно, было бы в высшей степени любопытно проанализировать, в какой степени параметр длительности в жестикуляции характерен для романских и германских языков.

4.2. Группа НСВ-глаголов

Как показывает табл. 2 и 7, параметр многократности важен для различения СВ- и НСВ-глаголов — СВ-глаголы сопровождаются преимущественно однократными жестами, а НСВ-глаголы — многократными. Содержательно интерпретировать этот факт позволяют данные в табл. 7. Мы видим, что для различения разных групп СВ-глаголов признак кратности несуществен, но он хорошо различает НСВ мультипликативы и континуативы. Тем самым, можно предположить, что разные виды многократных жестов (истинно многократные, квантующие, жестикуляционные повторы) с помощью повторяемых движений передают образ **гранулированных** мультипликативов (если пользоваться метафорами, предложенными в работах [Janda 2004, 2007, 2008]) в их противопоставлении **жидкостям** непрерывных процессов. В этой связи интересно вспомнить, что мультипликативы в ряде языков маркируются редупликацией (см. об этом [Плунгян, Рахилина 2007: 746] со ссылкой на [Плунгян 1989]); хорошо известный пример такой редупликации — др.-греч. κυκλοσ «колесо», в котором редупликации отражает итеративный процесс вращения колеса.

Таким образом, мы видим, что посредством многократности жеста в русской жестикуляции отражается факт, проницательно отмеченный в [Падучева 1994: 29]: «Если у глагола НСВ нет парного семельфактива, то нет и ощущения многоактности»: жестикуляция, как физическое отражение когнитивных процессов, лежащих в основе данного высказывания, отражает **потенциальную** возможность для данного глагола образовать семельфактив, хотя самого семельфактива в порождаемом высказывании нет.

Что касается группы СВ-глаголов, то они, будучи — в разной степени — мгновенными, нульмерными, точечными, краткими, не предполагают внутри себя дальнейшего расчленения на более мелкие элементы или кванты, как следствие, идея повтора и противопоставление однократных и многократных жестов для них попросту неактуальна.

4.3. Группа СВ-глаголов

Что касается СВ-глаголов, то, по нашим данным, внутри этой группы глаголы на *ану-* противопоставлены остальным бесприставочным СВ-глаголам,

в т. ч. и глаголам на *ну-* (и тут мы должны не согласиться с [Плунгян 2000: 217], где автор, в отличие от А. В. Исаченко и И. С. Улуханова, не «усматривает семантического различия между *ну-* и *ану-* по степени интенсивности»): жестикуляционные данные, а именно отчетливое тяготение жестикуляции, сопровождающей глаголы на *ану-*, к таким жестикуляционным событиям, как конфигурация *кулак* и движение типа *замах/отскок*, отчетливо демонстрирует, что с когнитивной точки зрения глаголы на *ану-* тесно связаны с идеей интенсивности, усилия, повышенной энергетики, которые передаются именно этими жестикуляционными компонентами.

Ситуация несколько напоминает ситуацию с русскими кванторами общности *весь* и *все*, которые — с лингвистической точки зрения — вполне можно считать одним и тем же словом (что исторически является абсолютной истиной). Однако в современном употреблении когнитивная картина мира, лежащая за этими словами и мотивирующая употребление сопровождающей жестикуляции, показывает, что для говорящего это, безусловно, разные слова (см. подробнее [Гришина 2015б]).

Действительно, согласно работе [Kuznetsova, Makarova 2012: 155], выбор между *ану-* и *ну-* в значительной степени (т. е. не абсолютно, а статистически) зависит от структуры основы — глаголы на *ану-* предпочитают односложную основу (*тормознуть* vs. **тормозануть*), т. е. выбор осуществляется под влиянием не семантических, а достаточно технических — в первую очередь, по-видимому, акцентологических — факторов (проблема требует специального исследования и аккуратного описания, здесь мы можем лишь упомянуть о ней). Ожидать в этом случае отражения различия между этими глаголами в жестикуляции довольно бессмысленно, поскольку морфонология и акцентология имеют весьма опосредованное отношение к когнитивным процессам, а следовательно, и к жестикуляции.

И тем не менее, на жестикуляционном уровне мы это различие наблюдаем.

Таким образом, согласно жестикуляционным данным, в глаголах с суффиксом *ану-* мы имеем дело с единственным, судя по всему, случаем выражения оценки с помощью глагольного суффикса. Структура значения в этом случае, очевидно, имеет двухуровневый характер, как и в существительных с оценочными суффиксами, например, *ищ(е)*. Можно сравнить: *кусище* ‘**кусок**, который **говорящий оценивает как большой**’, *кусануть* ‘совершить одноактный **акт кусания**, либо приложив **повышенные, с точки зрения говорящего, усилия,** либо **откусив слишком большой, с точки зрения говорящего, кусок**’ (при этом *куснуть* будет, очевидно, означать всего лишь ‘совершить одноактный акт кусания’). Именно этим, по-видимому, объясняются следующие факты.

1) Глаголы с *ану-*, в отличие от глаголов с *ну-*, как было отмечено в работе [Makarova 2009: 19], не сочетаются с приставками: *хлебать — хлебануть — *отхлебануть* vs. *хлебать — хлебнуть — отхлебнуть*. Приставки аспектуализируют процесс или событие, обозначаемые диктальной частью значения глагольной основы. Если же значение глагольной основы включает в себя, помимо диктальной, прагматическую (оценочную) зону, то аспектуализация последней

представляется невозможной, соответственно, невозможной оказывается и аспектуализация такой глагольной основы как семантического целого.

2) Глаголы с *ану-* иногда довольно легко образуются от НСВ-глаголов, которые не образуют глаголов на *ну-*, например, *резать* — *резануть*, но не **резнуть* (разумеется, такие образования, как *резануть*, относятся к периферийной зоне литературного языка, но, тем не менее, показателен сама возможность образования и однозначного понимания таких окказионализмов носителями русского языка). В случае **резнуть* мы сталкивается со стандартной ситуацией, когда исходный континуатив *резать* не может быть естественным образом разделен на минимальные порции или кванты[13]. Возможность же образования глагола *резануть* определяется тем, что в нем **краткость** ситуации переосмысливается как ее **однократность** (действие ‘резать’ осуществляется в один прием), а оценочная часть значения фиксирует, что данное однократное действие произведено субъектом с повышенной интенсивностью. Аналогично — *грабить/грабануть/*грабнуть, долбать/долбануть/*долбнуть, мешать* ‘перемешивать’/ *мешануть/*мешнуть, трусить* ‘бояться’/?*трусануть/*труснуть, цеплять/цепануть/*цепнуть, чесать/чесануть/?чеснуть*).
3) Глаголы на *ану-* с трудом образуются от исходных НСВ-глаголов, если у последних в составе значения уже есть элемент ‘сильно’ (*топать/*топануть/топнуть*): в этом случае семантический компонент ‘интенсивность’ уже входит в диктальную зону значения, что делает в принципе избыточной дополнительную прагматическую зону.
4) Глаголы с *ану-* с трудом образуются от НСВ-глаголов, у которых в составе значения есть компонент ‘слабо, мало’: *лизать/*лизануть/ лизнуть, кивать/*кивануть/кивнуть, икать/*иканути/икнуть, мелькать/*мелькануть/мелькнуть, мигать/*мигануть/мигнуть, моргать/*моргануть/моргнуть, порхать/*порхануть/порхнуть, трогать/*трогануть/тронуть, тыкать* (*пальцем*)/?*тыканути/ тыкнуть* или *ткнуть*, *фыркать/*фыркануть/фыркнуть*. И это естественно, поскольку в этом случае прагматическая зона значения (‘говорящий считает, что действие было произведено с интенсивностью выше нормы’) противоречит семантическому компоненту ‘действие производится с интенсивностью ниже нормы’.

[13] «В аспектологии же хорошо известно, что полноценная мультипликативная интерпретация существует только у обозначений таких ситуаций, единичный “квант” которых обладает достаточной прагматической значимостью для того, чтобы иметь в языке самостоятельное лексическое выражение» [Плунгян, Рахилина 2007: 746, со ссылкой на работы Е. В. Падучевой и В. С. Храковского]).

5. Заключение

Итак, можно подытожить, что русская жестикуляционная система имеет ряд способов отразить видовые значения в русском языке. При этом использование этих способов в значительной степени асимметрично.

Параметр многократности 1) не работает внутри группы СВ-глаголов (т. е. не различает моментальные глаголы на *ану-* и *ну-* и бесприставочные СВ-глаголы); 2) прекрасно работает в группе НСВ, различая мультипликативы, т. е. длительные потенциально членимые действия, и континуативы, т. н. длительные нечленимые действия; 3) поскольку этот параметр хорошо выражен в группе НСВ-глаголов и вовсе не выражен в группе СВ глаголов, то именно поэтому он различает СВ- и НСВ-глаголы, т. е. действие этого параметра на уровне различения видов — производно и зависит от различения разных типов НСВ глаголов.

Параметр энергетики, наоборот, 1) не работает внутри группы НСВ-глаголов, т. е. не различает мультипликативы и континуативы; 2) хорошо работает внутри СВ группы, различая интенсивные моментальные глаголы на *-ану-* и просто моментальные глаголы на *-ну-*; 3) и именно потому, что этот параметр хорошо работает в группе СВ и совсем не работает в группе НСВ, он и различает видовые пары СВ- и НСВ-глаголов, т. е. опять же, на уровне противопоставления видов он имеет вторичное значение.

Параметр длительности 1) не работает внутри группы СВ-глаголов; 2) не работает внутри группы НСВ-глаголов; 3) работает **только** в противопоставлении СВ- и НСВ-глаголов.

Из сказанного следует, что в русской жестикуляции только параметр длительности жеста передает именно **видовое** противопоставление СВ-НСВ, а не семантические особенности разных НСВ- и СВ-глаголов.

Добавим в завершение. Если принять, что разные жестикуляционные компоненты и параметры занимают разное место в системе русской жестикуляции, например, конфигурация ладони, направление движения руки или головы, траектория движения расположены ближе к **центру жестикуляционной системы** (т. е., с одной стороны, отражают существенные для понимания высказывания семантические и прагматические компоненты, а с другой, — легко воспринимаются и опознаются слушающим), а длительность жеста, его кратность, исполнение одной или двумя руками, «ручность», т. е. выбор между правой и левой рукой, степень напряжения руки и энергетика жеста, — ближе к **периферии** жестикуляционной системы, то придется признать, что видовые противопоставления являются периферийными для русской жестикуляции. В этой связи было бы в высшей степени полезным и любопытным сравнить наши данные с аналогичными данными для других языков, а также проанализировать, уже полностью отвлекшись от чисто видовых противопоставлений, жестикуляционное отражение таксономической системы русских глаголов. Но это — предмет отдельного, довольно большого и трудоемкого исследования, имея в виду, к тому же, что на результаты будут оказывать влияние не только жестикуляционный профиль русского вида, но и жестикуляционные профили приставок и корней.

Литература

1. *Bressem, Ladewig 2011* — Bressem J., Ladewig S. H. Rethinking gesture phases: Articulatory features of gestural movement? // Semiotica. № 184. 2011. P. 53–91.
2. *Croft 2012* — Croft, W. (2012). Verbs: Aspect and causal structure. Oxford University Press.
3. *Janda, Laura A.* (2004) A metaphor in search of a source domain: the categories of Slavic aspect. Cognitive Linguistics 15(4), pp. 471–527.
4. *Janda, Laura A.* (2007) Aspectual clusters of Russian verbs. Studies in Language 31(3), pp. 607–648.
5. *Janda, Laura A.* (2008) Semantic Motivations for Aspectual Clusters of Russian Verbs. In Christina Y. Bethin (ed.), American Contributions to the 14th International Congress of Slavists, Ohrid, September 2008. Bloomington: Slavica Publishers, pp. 181–196.
6. *Kendon 1972* — Kendon A. Some relationships between body motion and speech // A. Siegman, B. Pope Studies in dyadic communication New York, 1972. P. 177–210.
7. *Kendon 1980* — Kendon A. Gesticulation and speech: Two aspects of the process of utterance // M. R. Key The relation between verbal and nonverbal communication The Hague, 1980. P. 207–227.
8. *Kita et al. 1998* — Kita S., Van Gijn I., Van der Hulst H. Movement phases in signs and co-speech gestures, and their transcription by human coders // Gesture and sign language in human-computer interaction Berlin Heidelberg, 1998. P.23–35.
9. *Kuznetsova, Makarova 2012* — Julia Kuznetsova and Anastasia Makarova. distribution of two semelfactives in russian: -nu- and -anu- // A. Grønn & A. Pazelskaya (eds.) The Russian Verb, Oslo Studies in Language 4(1), 2012. 155–176.
10. *László 2008* — Jászay László. Концепт моментальности в системе русского глагола. Slavica Szegediensia VI. Szeged, 2007–2008 (88–96).
11. *Makarova 2009* — А. Макарова. Психолингвистические данные об алломорфии в русских семельфактивах, Tromsø, 2009.
12. *McNeill et al. 2001* — McNeill et al. 2001 — McNeill D., Quek F., McCullough K.-E., Duncan S., Furuyama N., Bryll R., Ma X.-F., Ansari R. Catchments, Prosody, and Discourse // Gesture. V. 1. № 1. 2001. P. 9–33.
13. *Гришина 2013* — Гришина Е. А. Жестикуляционные профили русских приставок // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая — 2 июня 2013 г.). Вып. 12 (19). — М.: Изд-во РГГУ, 2013, с. 255–271.
14. *Гришина 2014* — Гришина Е. А. Жесты и прагматические компоненты высказывания // Мультимодальная коммуникация: теоретические и эмпирические исследования. Под ред. О. В. Федоровй и А. А. Кибрика. М., «Буки Веди», 2014, с. 25–47.
15. *Гришина 2015а* — Гришина Е. А. Круги и колебания: семантика сложных траекторий в русской жестикуляции // Язык и мысль: современная когнитивная лингвистика / Сост. А. А. Кибрик, А. Д. Кошелев; ред. А. А. Кибрик и др. — М.: Языки славянской культуры, 2015. — 848 с., илл. — (Вклейка после с. 368.) — (Разумное поведение и язык. Language and Reasoning). с. 238–286.

16. *Гришина 2015б* — Гришина Е. А. Кванторные слова, жестикуляция и точка зрения // Компьютерная лингвистика и интеллектуальные технологии По материалам ежегодной Международной конференции «Диалог». Вып. 14. 2015. С. 181–198.
17. *Падучева 1994* — Е. В. Падучева. Таксономические категории глагола и семантика видового противопоставления. Семиотика и информатика, вып. 34, М., 1994, с. 7–31.
18. *Плунгян 2000* — В. А. Плунгян. ‘Быстро’ в грамматике русского и других языков // Л. Л. Иомдин, Л. П. Крысин (ред.), Слово в тексте и в словаре: Сб. статей к 70-летию акад. Ю. Д. Апресяна. М.: ЯРК, 2000, 212–223.
19. *Плунгян В. А.* Выражение множественности ситуаций в чамалинском языке // Храковский В. С. (ред.). Типология итеративных конструкций. Л.: Наука, 1989. С. 79–87.
20. *Плунгян, Рахилина 2007* — В. А. Плунгян, Е. В. Рахилина. К типологии глаголов ‘летать’ и ‘прыгать’ // Т. А. Майсак, Е. В. Рахилина (ред.). Глаголы движения в воде: лексическая типология. М.: Индрик, 2007, 739–748.
21. *Янда 2015* — Янда, Л. А. (2015). Аспектуальные типы русского глагола: пересматривая типологию Крофта. // Язык и мысль: современная когнитивная лингвистика / Сост. А. А. Кибрик, А. Д. Кошелев; ред. А. А. Кибрик и др. — М.: Языки славянской культуры, 2015. — 848 с., илл. — (Вклейка после с. 368.) — (Разумное поведение и язык. Language and Reasoning). 213–237.

СТРУКТУРА ДВУХМЕСТНЫХ КОННЕКТОРОВ РУССКОГО ЯЗЫКА В СВЕТЕ КОРПУСНЫХ ДАННЫХ[1]

Инькова О. Ю. (Olga.Inkova@unige.ch)
Женевский университет, Женева, Швейцария; Институт проблем информатики ФИЦ ИУ РАН, Москва, Россия

Попкова Н. А. (Natasha__popkova@mail.ru)
Институт проблем информатики ФИЦ ИУ РАН, Москва, Россия

В работе рассматривается проблема вариативности формы двухместных коннекторов русского языка на примере *не то чтобы... (но)* и *не то что... (а)*. Проведенный с использованием формального и функционально-семантического критериев анализ позволяет говорить о разных языковых единицах, первая из которых выражает отношение замещения по степени дескриптивной достоверности, вторая, помимо этого отношения, выражает еще и отношение замещения по большей степени аргументативной значимости, что связано с разной сферой действия входящего в их состав отрицания *не*, тогда как направление градации в обоих отношениях одинаково (восходящее). Поскольку *не то что* и *не то чтобы* могут устанавливать соответствующие отношения самостоятельно, то их можно считать исходной — минимальной — формой показателей данных отношений. Их употребление в составе двухместных коннекторов в сочетании с противительными союзами *но* и *а* и другими лексемами, совместимыми с семантикой устанавливаемого отношения, диктуется коммуникативным замыслом говорящего, выбранной синтаксической конструкцией и др. дискурсивными параметрами и должны квалифицироваться как речевые реализации показателей соответствующих отношений. Предлагаемые решения подтверждаются данными НКРЯ.

Ключевые слова: коннекторы, семантика, вариативность, русский язык, *не то чтобы, не то что*

[1] Работа выполнена при поддержке РФФИ (грант № 16-06-00070), РГНФ (грант № 16-24-41002) и ШННФ / FNS (грант № IZLRZ_164059/1).

THE STRUCTURE OF TWO-PART CORRELATIVE CONNECTORS AS AN OBJECT OF CORPUS ANALYSIS

Inkova O. Yu. (Olga.Inkova@unige.ch)
University of Geneva, Geneva, Switzerland; Institute of Informatics Problems, FRC CSC RAS, Moscow, Russia

Popkova N. A. (Natasha__popkova@mail.ru)
Institute of Informatics Problems, FRC CSC RAS, Moscow, Russia

The paper discusses the problem of formal variability of Russian two-part correlative connectors on the example of *ne to chtoby…no* and *ne to chto…a*. The results of the analysis, carried out both with formal and functional-semantic criteria, allow to state that *ne to chtoby… no* and *ne to chto… a* are two separate linguistic units with the first expressing substitution aimed towards more descriptive adequacy and the second unit expressing, beyond that, substitution aimed towards more argumentative relevance. This semantic difference is due to the different scope of the negative particle *ne*, which is the part of both markers; even if in both cases the gradation is rising. The position of *ne to chtoby* and *ne to chto* is not fixed, and *ne to chtoby* can be characterized by phonetic (*ne to chtob*) and morphological (*ne tak chtob(y)*) variability. As forms *ne to chtoby* and *ne to chto* can express relations of substitution alone, they may be considered basic or minimal markers of such relations. The use of these forms as two-part correlative connectors with adversative conjunctions *no, a* and other lexical units is dictated by the speaker's communicative intention, the syntactical construction and other discursive parameters. The Russian National Corpus data confirms our statements.

Keywords: connectors, semantics, variability, Russian, *ne to chtoby, ne to chto*

Введение

Структура коннекторов русского языка (да и не только русского; см. дискуссию в Fraser 2013) является мало разработанным вопросом, хотя он был поставлен еще в статье [Прияткина 1977]. В настоящее время, в связи с созданием аннотированных корпусов и надкорпусных баз данных, в том числе коннекторов [см. Инькова, Кружков 2016], этот вопрос становится особенно актуальным. Прежде всего, необходимо определить линейные границы языковых единиц, состоящих из более чем одного 'слова' (иначе говоря, из компонентов, разделенных пробелом и/или другими словами). Во-вторых, необходимо определить исходную форму языковых единиц, характеризующихся вариативностью.

Эти два свойства — многокомпонентность и вариативность — присущи многим коннекторам русского языка (ср. *но — но однако еще и, ...; да — да к тому же; да притом, да еще, ...; не только… но/ но даже/ а также/…*).

В этой связи возникает вопрос о критериях, позволяющих определить, когда можно говорить о разных языковых единицах, а когда о вариантах одной, и какой из вариантов можно считать базовым. В [Инькова 2016], на примере двухместного коннектора *не только… но*, данную проблему было предложено решать исходя из двух теоретических постулатов. Прежде всего, необходимо развести концептуальный и языковой уровень анализа. В случае коннекторов это, соответственно, тип выражаемого логико-семантического отношения (противительные, временные, уступительные и др.) и языковые средства, которыми располагает язык для его выражения (так, уступительные отношения могут выражаться в русском языке коннекторами *хотя, тем не менее, однако* и др.). Кроме того, предлагается различать единицы, принадлежащие системе языка (их выбор диктуется правилами грамматики), и единицы, принадлежащие системе речи (их выбор определяется коммуникативным замыслом говорящего). Преимущества предлагаемого подхода заключаются в том, что при анализе формы коннекторов и ее вариативности он не исходит из априорного существования в системе языка одной единицы с бесконечным количеством вариантов, а из того, что говорящий часто, исходя из динамики текста или коммуникативной ситуации, буквально «по кирпичикам» (которые сами могут быть уже многокомпонентными) собирает необходимый для его замысла показатель логико-семантического отношения.

В настоящем исследовании данный подход будет применен к коннекторам *не то что…, не то… чтобы… (а, но)* — так они представлены в БАС — с целью сформулировать критерии, позволяющие определить идет ли речь об одной языковой единице, как предлагают некоторые, или о разных. Анализ опирается на данные Национального корпуса русского языка (www.ruscorpora.ru).

1. История вопроса

В русской грамматической традиции нет единой точки зрения ни относительно формы интересующих нас коннекторов, ни относительно их семантики, ни относительно их количества. Что касается формы, то ее варьирование может касаться следующих параметров:

- количество компонентов коннектора: второй фрагмент текста может быть немаркирован:

(1) *Кавторанг и адмирал* ***не то чтобы*** *беседовали, что-то обсуждали или о чём-то спорили — они выносили приговор без права обжалования.*
[Юрий Давыдов. Синие тюльпаны (1988–1989)]

- форма первой части коннектора: элемент *бы* может сокращаться до *б* (2) или отсутствовать (3), а демонстратив *то* может замещаться *так* (ср. (9) ниже):

(2) *Бродя туда-сюда по узким тропкам меж особняков, Шабашов* ***не то чтоб*** *мерз,* ***но*** *чувствовал озноб.* [Андрей Дмитриев. Призрак театра (2002–2003)]

(3) *С этих пор и завязалось у него с книголюбом* ***не то что*** *приятельство,* ***а*** *хорошее знакомство.* [Ю. О. Домбровский. Ручка, ножка, огуречик (1977)]

- вторая часть коннектора: вместо или вместе с *но* может быть использована другая лексема противительной семантики: *а* (3), *напротив* (4)*, просто* (6) и (13) и др.

(4) ***Не то чтобы*** *нас это удивляет.* ***Напротив****, весьма радует.* [Юлия Пешкова. Дом мод (2002)]

- порядок следования фрагментов текста: фрагмент, содержащий компонент *не то что(бы)* может занимать как первую, так и вторую позицию (5) и (8) ниже:

(5) *Но я продолжал стоять, потому что чувствовал — мне и стоя трудно будет их убедить,* ***не то что*** *лежа.* [Фазиль Искандер. Время счастливых находок (1973)]

Существует, как минимум, четыре точки зрения относительно количества языковых единиц с компонентами *не то что* и *не то чтобы*. Согласно первой, наименее распространенной, существует один союз *не то чтобы… но и* [Завьялов 2009: 27, 120, 194 сл.], характеризующийся формальной вариативностью. Согласно второй точке зрения, прямо противоположной и представленной, в основном, в словарях служебных слов, каждый формальный вариант может претендовать на статус самостоятельной языковой единицы [Рогожникова 2003, Ефремова 2004, Леденев 2006, Цой 2008, Бурцева 2010]. Например, [Цой 2008] приводит список из одиннадцати единиц.

Сторонники третьей точки зрения полагают, что существуют два двухместных союза (с наиболее часто фиксируемыми формами *не то чтобы… но/а* и *не то что… а*). Некоторые из них [Рогожникова 1971, Санников 2008, Черемисина, Колосова 2009: 143] признают за обоими союзами или только за *не то что… а* возможность как изменения порядка следования компонентов, так и самостоятельного употребления первой части союза [Серебряная 1972: 91, Ляпон 1986, Инькова-Манзотти 2001: 334][2]. Наконец, согласно четвертой точке зрения [Ушаков, Шведова 1960: 236, Морковкин 1997], существует частица *не то что(бы)*, которая может выступать также в составе двухместного союзного соединения.

---

[2] [РГ-80] придерживается неоднозначной позиции: признается существование двух союзов с вариативной второй частью: *не то чтобы… но* и *не то что… а* [РГ-80, §1681, 2078], но в «Предметном указателе» [РГ-80 : II, 683] фигурируют шесть языковых единиц.

Что касается семантики интересующих нас коннекторов, то, за исключением [Пешковский 2001: 432, БАС], где *не то чтобы* и *не то что* относятся к союзам пояснительным, их считают градационными. Однако и здесь нет единства. Так, сторонники первой точки зрения, признающей существование единой языковой единицы — частицы *не то что(бы)* или двухместного союза с вариативными частями — приписывают им либо два значения (1. для уточнения при отрицании возможного предположения или при поправке к сказанному и т. п.; 2. в знач. *не только* [Ушаков]), либо только первое из них [Шведова 1960: 236, Морковкин 1997]. Это утверждение противоречит, однако, корпусным данным. Не соответствует языковой действительности и утверждение, согласно которому *не то чтобы* и *не то что*, две самостоятельные языковые единицы, различаются направлением градации: нисходящая для первого, восходящая для второго [ср. Серебряная 1972, Санников 2008]. В данном подходе смешиваются, как будет показано, разные типы градационных шкал.

Наконец, [Рогожникова 1971, РГ-80: §3154, 3155, Инькова-Манзотти 2001: 338] признают у коннектора с компонентом *не то чтобы* одно значение — отрицания описания некоторого положения вещей и замещение его более достоверным, а у коннектора с компонентом *не то что*, помимо этого значения, признается еще одно: выражение большей аргументативной значимости второго фрагмента текста, близкое к значению коннекторов с компонентом *не только*[3]. Иными словами, во втором значении мы как будто имеем дело с отношением не замещения, а добавления.

Возникает вопрос, который, однако, до сих пор не был поставлен: каким образом одна и та же языковая единица может выражать два противоположных значения? Объяснение, по-видимому, можно найти во взаимодействии отрицания *не* (и его сферы действия) с механизмом градации, лежащим в основе отношений, устанавливаемых коннекторами с начальными компонентами *не то чтобы* и *не то что*, а также в семантической природе фрагментов текста, между которыми устанавливается отношение.

2. Семантика «не то чтобы»

Коннекторы с компонентом *не то чтобы* устанавливают отношение замещения менее достоверного описания (фрагмент текста, маркируемый *не то чтобы*) более достоверным описанием того же положения вещей. Так, в (1) выше речь идет о возможных описаниях того, как общались Кавторанг и адмирал. Описание *беседовали, что-то обсуждали или о чём-то спорили* (*q*)

---

[3] В работе [Инькова-Манзотти 2001: 336] у коннекторов с начальным компонентом *не то что* и *не то чтобы* выделяется еще одно значение: «отрицание возможного объяснения» или отношение мотивации. Но это отношение не между фрагментами текста, соединяемыми данными коннекторами, а отношение между высказыванием, оформляемым данными коннекторами, и предыдущим высказыванием. Поэтому оно может считаться текстообразующей функцией данных коннекторов, а не одним из их значений [ср. Зайцева 1998].

замещается более соответствующим истине: *они выносили приговор без права обжалования* (*p*), иначе: ¬ *q* ∧ *p*, где «¬» — отрицание, а *q* и *p* — два описания одного положения вещей, сопоставляемые на шкале истинности.

В таких высказываниях отрицание в составе *не то чтобы* имеет полемический характер. Действительно, отрицаемое в *q* описание часто эксплицитно присутствует в предыдущем контексте: это либо точка зрения другого говорящего, потенциального или реального (6), либо самого говорящего, как бы строящего внутренний диалог в поисках точного описания; ср. (2) выше.

(6) *— Вы так его не любите? Он поморщился. — Да нет,* ***не то чтобы*** *я не люблю его,* ***но просто****…* [Ю. О. Домбровский. Ручка, ножка, огуречик (1977)]

Отрицание в *не то чтобы*, кроме того, имеет сентенциальную сферу действия, как синтаксическую, так и семантическую. Доказательством тому — частая парцелляция второго компонента (см. (4) выше), отсутствие синтаксической симметрии соединяемых фрагментов текста, преимущественно начальная (или после общих для *q* и *p* элементов) позиция *не то чтобы*, употребление отрицательной частицы *нет* после первого фрагмента текста (7), а в диалоге — до него (6), что свидетельствует также и о том, что речь идет именно об отрицании и о замещении компонента, маркируемого *не то чтобы* [Инькова-Манзотти 2001: 338].

(7) *И* ***не то чтобы*** *за ним числилось нечто неблаговидное или криминальное.* ***Нет. Просто*** *был у нас начальником отдела кадров Николай Александрович Махов, который не любил евреев в принципе, а Фрадкиса вдвойне.* [И. Э. Кио. Иллюзии без иллюзий (1995–1999)]

Градация производится на шкале достоверности: отрицаемое описание *q* занимает на этой шкале более низкое положение (оно менее достоверно), чем утверждаемое *p*; следовательно, направление градации — восходящее. Другое дело, что довольно часто, но не всегда (ср. (7) выше), шкала достоверности опирается на другую шкалу — шкалу некоторого признака [Инькова 2013]. Так, в (2) выше *q* = *мерз* и *p* = *чувствовал озноб* можно расположить на шкале «кому-то холодно», на которой *q* занимает более высокое положение, чем *p*, и в силу этого отвергается как обозначающее слишком высокую степень проявления данного признака. Ср. также (8) ниже: отрицаемое *q* = *весело до упаду* занимает на шкале «хорошо провести время» более высокое положение, чем *p* = *вполне мило*. По-видимому, именно это дает возможность исследователям говорить о нисходящей градации в высказываниях с *не то чтобы*.

Отношение замещения по степени достоверности может быть установлено *не то чтобы* и при обратном следовании фрагментов текста *q* и *p*: *p, не то чтобы q*, но при поддержке противительного союза, вводящего *q*:

(8) *Тоже сидели шесть часов, вполне мило,* ***но не то чтобы*** *весело до упаду.* [Василий Катанян. Лоскутное одеяло (1990–1999)]

Не то чтобы в таких высказываниях может считаться «позиционным вариантом» [Приняткина 1977] *не то чтобы* в высказываниях с *не то чтобы q, p*.

Что касается форм *не то чтоб* (2), *не так чтоб(ы)*, то они устанавливают между соединяемыми фрагментами текста тот же тип отношения: замещение по степени достоверности описания. Ср. (9), где хорошо виден полемический характер отрицания:

(9) *— Крадёт, что ли? —* ***Не так чтобы*** *крадёт,* ***но*** *пользуется.* [И. Грекова. Дамский мастер (1963)]

На этом основании их можно рассматривать как фонетический (*не то чтоб*) и морфологический (*не так чтобы* с его фонетическим вариантом *не так чтоб*) варианты *не то чтобы*, характеризующиеся разговорной и стилистической окраской, а также меньшей частотностью: НКРЯ фиксирует 538 вхождений для *не то чтоб*, 81 для *не так чтоб*, 265 для *не так чтобы* и 3206 для *не то чтобы*.

3. Семантика «не то что»

Анализ семантики *не то что* мы начнем со значения, в котором он противопоставляется *не то чтобы* и которое можно определить как установление большей аргументативной значимости второго компонента.

(10) *Такого приказа* ***не то что*** *царь,* ***но и*** *татары,* ***и*** *немецкие оккупанты не подписывали.* [Василий Гроссман. Все течет (1955–1963)]

Градация производится на аргументативной шкале, ее направление — восходящее: чтобы убедить слушающего, говорящий замещает менее сильный аргумент более сильным. Так, говоря о жестокости приказа, говорящий увеличивает силу своих аргументов: *царь — татары — немецкие оккупанты*, располагая эти элементы на шкале жестокости по степени возрастания. Таким образом, здесь шкала аргументативной силы также опирается на шкалу признака, но их направления совпадают. Кроме того, отношение замещения здесь устанавливается не между двумя номинациями одного положения вещей, а между двумя положениями вещей.

Отрицание сохраняет свой полемический характер, но сфера его синтаксического действия — присловная: она распространяется на элемент, перед которым стоит *не то что*[4]. Об этом говорит тот факт, что модифицируемый *не то что* элемент, может находиться в сфере действия отрицания предиката

[4] О продуктивности различия между присловным и фразовым отрицаниями, а также о несовпадении сферы действия – синтаксической и семантической — отрицания см., например, [Падучева 2013, глава 5], других языковых единиц — [Guimier 1996: 1–7].

(11), в отличие от *не то чтобы*, которое распространяет свою сферу действия и на предикатное отрицание (12):

(11) ***Не то что*** *на сына миллионера, он даже на городского не похож.* [Сергей Юрьенен. Покер с Ильичом (1997)]

(12) ***Не то чтобы*** *Ольга их* ***не*** *приглашала.* ***Приглашала****. Но те, что приходили* (…) [Вера Белоусова. Второй выстрел (2000)]

При этом отрицание предиката изменяет направление градации шкалы признака: в (11) на шкале «элегантности» *сын миллионера* очевидным образом занимает более высокое положение, чем *городской*. Но *не походить на сына миллионера* является менее сильным аргументом в пользу неэлегантности, чем *не походить даже на городского*. Об этом же говорит употребление *даже*. С семантической же точки зрения, отрицание в *не то что* является внешним по отношению к пропозиции, в которой он находится, не затрагивая ее истинности; иными словами, (11) выше не означает *Неверно, что он не похож на сына миллионера, верно, что он не похож на городского*. Отрицается лишь достаточность *q* как аргумента в пользу некоторого вывода *r*: ¬ Arg (*q, r*) ∧ Arg (*p, r*), где «¬» — отрицание, «Arg (..., ...) » — аргументативная функция, *q* и *p* — два положения вещей, сопоставляемые на шкале аргументативной силы.

Отрицание менее сильного и утверждение более сильного аргумента приводит в некоторых случаях к семантическому эффекту добавления: отрицаемое положение вещей обозначает недостаточную степень проявления признака, поэтому оно как бы включается в утверждаемое положение вещей, занимающее на этой шкале более высокое положение. Отсюда часто отмечаемая исследователями семантическая близость *не то что* и *не только*. Ср. возможность их мены в (10)–(11). Однако, если *не только* указывает, что маркируемый им элемент принадлежит ко множеству ему подобных [Инькова 2016], то *не то что* сохраняет свою заместительную семантику. Ср. различие в интерпретации (13), где *не вписывались* замещается на *не помещались*, и его трансформацию с *не только*, включающим оба предиката во множество:

(13) (…) *старинная мебель или другие вещи, рассчитанные на просторные комнаты с высокими потолками,* ***не то что*** *не вписывались* ***—*** *они просто не помещались в малогабаритных квартирах* (…). [И. К. Архипова. Музыка жизни (1996)]

(13’)… ***не только*** *не вписывались — они просто не помещались.*

Поэтому такая мена возможна лишь в тех случаях, когда осуществление *p* и *q* не взаимоисключают друг друга. Ср. неприемлемость *не только* в (14):

(14) ***Не то что*** *отбой, по-моему уже гимн слышно откуда-то.* [Александр Солженицын. В круге первом (1968)] ≠ *не только*

Отношение большей аргументативной значимости может быть установлено и при обратном порядке следования *q* и *p*, причем *не то что* может самостоятельно выполнять связующую функцию; ср. (5) выше. Поскольку отношение между *q* и *p* остается тем же, *не то что* в высказываниях с порядком *p, не то что q* может рассматриваться как позиционный вариант *не то что* в высказываниях с *не то что q, p*.

Однако *не то что* может устанавливать отношение замещения не только на аргументативной шкале, но и на шкале достоверности описания. Это происходит благодаря расширению синтаксической сферы действия отрицания до сентенциальной[5]: в его сферу попадает и предикат, как у *не то чтобы*, а соединяемые фрагменты текста должны восприниматься как два возможных описания одного и того же положения вещей; ср. (15), а также (3) выше, где мена *не то что* на *не то чтобы* возможна, в отличие, например, от (10), (11) или (14):

(15) *И опять-таки* ***не то что*** *не было уж решительно никакого выхода немецкому народу, — выход был, и такой, и эдакий, — но стало ясно, что всё пошло прахом*. [Ю. О. Домбровский. Обезьяна приходит за своим черепом. Пролог (1943–1958)]

Принято считать, что *не то что* в этом значении более экспрессивен, тогда как *не то чтобы* вводит более смягченное отрицание первого компонента в силу «актуализации предположительного характера сообщаемого в соответствующей части» [Ляпон 1986: 98].

4. Семантика языковых средств, вводящих второй фрагмент текста (*p*)

Наши выводы о семантике *не то чтобы* и *не то что* подтверждаются и тем набором языковых средств, которые используются для введения второго фрагмента текста (*p*), который, впрочем, может быть никак не маркирован; ср. (1) выше. Но это не меняет значения высказывания: полемическое отрицание в составе *не то что /не то чтобы* создает ожидание продолжения [ср. Серебряная 1972: 92], и логико-семантическое отношение между *q* и *p* остается тем же. Кроме того, не все высказывания с немаркированным *p* можно легко преобразовать в высказывания с двухместным коннектором [Сигал 2004: 143–144]. Это позволяет заключить, что показателями отношения замещения по степени дескриптивной достоверности или аргументативной значимости являются именно *не то чтобы* и *не то что*. Именно их следует считать базовой, минимальной, формой показателя соответствующих отношений.

---

[5] Такое расширение сферы действия отрицания характерно и для других коннекторов, опирающихся на механизм градации: например, *не столько… сколько, насколько… настолько*; см. [Инькова 2013].

Если же говорящий выбирает маркировать второй компонент каким-то языковым средством, то однозначно предсказать его выбор невозможно. Можно лишь определить семантику лексем, способных появляться во втором фрагменте текста, на основе их совместимости с семантикой устанавливаемого *не то что* и *не то чтобы* отношения. Так, в высказываниях с отношением дескриптивной достоверности второй компонент может вводиться союзами *но* и *а*, часто в сочетании с лексемами, позволяющими выразить колебания или предпочтение говорящего в выборе словесной формы (*скорее, просто, как бы, что ли* и др.), с показателями контраста (*наоборот, напротив*), отражающими противительный характер отношения, или с лексемами, имеющими градационное значение (*уже, даже*). Наоборот, частица *и* аналогии не будет употребляется в таких высказываниях (ср. (2), (3) выше), в отличие от высказываний с отношением аргументативной значимости, где, как мы видели, может возникать семантический эффект добавления второго аргумента (ср. (10) выше).

В этой связи хотелось бы кратко остановиться на высказываниях с отношением дескриптивной достоверности, где второй компонент тем не менее вводится *но и*.

(16) *Она чуть покашивала светлым глазом, потряхивала седоватой гривой, не то чтоб порицая,* ***но и*** *не присоединяясь, а так, значит, смиряясь.*
[Светлана Васильева. Триптих с тремя неизвестными (2001)]

Необходимым условием для появления частицы *и* аналогии является наличие в *p* отрицания или лексемы с отрицательной полярностью. Замещающее описание признается таким образом тоже недостоверным и замещается третьим (в нашем примере *q = не порицая, p = не присоединяясь* оба замещаются *s = смиряясь*). Этим и объясняется появление частицы *и*, выражающей аналогию отрицания и которая иначе, в силу заместительной семантики отношения дескриптивной достоверности, была бы невозможна.

5. Выводы

Проведенный анализ позволяет сделать вывод о том, что *не то чтобы* и *не то что* являются разными языковыми единицами. Несмотря на свою формальную близость, они не тождественны по значению: первый из них выражает отношение замещения по степени дескриптивной достоверности, второй — помимо этого отношения, еще и отношение замещения по большей степени аргументативной значимости с соответствующими особенностями функционирования. При этом *не то что* и *не то чтобы* имеют позиционные варианты, а *не то чтобы* характеризуется также фонетическим — *не то чтоб* и морфологическим — *не так чтоб(ы)* — варьированием. Именно комбинация двух критериев — формального и функционально-семантического — позволяет определить, идет ли речь об одной и той же языковой единице или нет. Например, коннекторы *а то* и *а так*, несмотря на аналогичное морфологическое варьирование

(мена демонстратива *то / так*), должны считаться разными языковыми единицами, так как они выражают разные логико-семантические отношения.

Исходными формами показателей отношений замещения по большей достоверности или аргументативной силе должны считаться *не то чтобы* и *не то что*, поскольку они могут устанавливать эти отношения самостоятельно. Их употребление в составе двухместных коннекторов в сочетании с противительными союзами *но* и *а* или другими лексемами, совместимыми с семантикой устанавливаемого отношения, диктуется коммуникативным замыслом говорящего, выбранной синтаксической конструкцией и др. дискурсивными параметрами и должны быть квалифицированы как речевые реализации показателей соответствующих отношений. Это подтверждает выводы, сделанные в [Инькова 2016] на примере *не только… но и*, и позволяет говорить об определенной закономерности.

Литература

1. *БАС* — Словарь современного русского литературного языка: В 17 т., Москва — Ленинград, АН СССР.
2. *Бурцева В. В. сост.* (2010), Словарь наречий и служебных слов русского языка, Москва, Русский языка–Медия; Дрофа.
3. *Ефремова Т. Ф.* (2004), Толковый словарь служебных частей речи русского языка, Москва, Астрель-Аст.
4. *Завьялов В. Н.* (2009), Морфологические и синтаксические аспекты описания структуры союзов в современном русском языке. Дисс. доктора филол. наук, Владивосток, ДГУ.
5. *Зайцева Г. Д.* (1998), Семантико-синтаксические свойства союза *не то чтобы…, но/а*, Семантика языковых единиц. Доклады VI Международной конференции, Москва, СпортАкадемПресс, сс. 101–103.
6. *Инькова-Манзотти О. Ю.* (2001), Коннекторы противопоставления во французском и русском языках. Сопоставительное исследование, Москва, Информэлектро.
7. *Инькова О. Ю.* (2013), О семантике так называемых градационных союзов *не столько… сколько* и *скорее… чем* // Вопросы языкознания, № 1, сс. 38–52.
8. *Инькова О. Ю.* (2016), К проблеме описания многокомпонентных коннекторов русского языка: *не только… но и* // Вопросы языкознания, № 2, сс. 37–60.
9. *Инькова О. Ю., Кружков М. Г.* (2016), Надкорпусные русско-французские базы данных глагольных форм и коннекторов // Linguistica e filologia, в печати.
10. *Леденев Ю. И.* (2006), Словник словаря неполнозначных слов русского языка, Ставрополь, СГУ.
11. *Ляпон М. В.* (1986), Смысловая структура сложного предложения и текст: К типологии внутритекстовых отношений, Москва, Наука.

12. *Морковкин В. В. ред.* (1997), Словарь структурных слов русского языка, Москва, Лазурь.
13. *Падучева Е. В.* (2013), Русское отрицательное предложение, Москва, Языки славянской культуры.
14. *Пешковский А. М.* (2001), Русский синтаксис в научном освещении, Москва, УРСС.
15. *Прияткина А. Ф.* (1977), Об отличии союза от других связующих слов // Русский язык в школе, № 4, сс. 102–106.
16. *РГ-80* — Русская грамматика / Под ред. Н. Ю. Шведовой, Москва, Наука, 1980.
17. *Рогожникова Р. П.* (1971), Градационные союзы в русском языке // Русский язык в школе, № 3, сс. 84–89.
18. *Рогожникова Р. П.* (2003), Толковый словарь сочетаний, эквивалентных слову, Москва, Астрель-АСТ.
19. *Санников В. З.* (2008), Русский синтаксис в формально-прагматическом аспекте, Москва, Языки русской культуры.
20. *Серебряная Ф. И.* (1972), К вопросу о структуре градационного ряда // Русский язык в школе, № 2, сс. 89–93.
21. *Сигал К. Я.* (2004), Сочинительные конструкции в тексте: опыт теоретико-экспериментального исследования (на материале простого предложения). Дис. доктора филол. Наук, Москва, ИЯ РАН.
22. *Ушаков Д. Н. ред.* (2001), Толковый словарь русского языка, Москва, Вече.
23. *Цой А. С.* (2008), Служебные слова как объект русской лексикографии, Москва, Издательство Литературного института.
24. *Черемисина М. И., Колосова Т. А.* (2010), Очерки по теории сложного предложения, Москва, УРСС (1-е изд. Новосибирск: Наука, 1987).
25. *Шведова Н. Ю.* (1960), Очерки по синтаксису русской разговорной речи, Москва, АН СССР.

References

1. *BAS* — The dictionary of modern standard Russian: in 17 volumes [Slovar' sovremennogo russkogo literaturnogo jazyka: v 17 t.], Moscow, Leningrad, Academy of Sciences of the USSR.
2. *Burtseva V. V.* (2010), Dictionary of Russian adverbs and auxiliary words [Slovar' narechij i sluzhebnyh slov russkogo jazyka], Moscow, Russkij jazyk–Media, Drofa.
3. *Cheremisina M. I., Kolosova T. A.* (2010), An outline of the complex sentence theory [Ocherki po teorii slozhnogo predlozheniya], Moscow, URSS (1st ed. Novosibirsk, Nauka, 1987).
4. *Efremova T. F.* (2004), Explanatory dictionary of the Russian auxiliary parts of speech [Tolkovyi slovar' sluzhebnykh chastei rechi russkogo yazyka], Moscow, Astrel'-Ast.

5. *Fraser B.* (2013), Combinations of Contrastive Discourse Markers in English, International Review of Pragmatics, No. 5, pp. 318–340.
6. *Guimier C.* (1996), Adverbs in French : the case of the adverbs in -ment [Les adverbes du français: le cas des adverbes en -ment], Paris, Ophrys.
7. *In'kova-Manzotti O. Yu.* (2001), Connectors of opposition in French and Russian. A comparative study [Konnektory protivopostavleniya vo frantsuzskom i russkom yazykakh. Sopostavitel'noe issledovanie], Moscow, Informelektro.
8. *In'kova O. Yu.* (2013), On the semantics of the so called gradation conjunctions *ne stol'ko… skol'ko* and *skoree… chem* [O semantike tak nazyvaemykh gradatsionnykh soyuzov *ne stol'ko… skol'ko* i *skoree… chem*], Voprosy yazykoznanija, No. 1, pp. 38–52.
9. *Inkova O. Yu.* (2016), Towards the description of multiword connectives in Russian: *ne tol'ko… no i* (not only… but also) [K probleme opisaniya mnogokomponentnykh konnektorov russkogo yazyka: *ne tol'ko… no i*], Voprosy yazykoznaniya, No. 2, pp. 37–60.
10. *Inkova O., Kruzhkov M.* (2016), Supracorpora databases of Russian and French verbal forms and connectors [Nadkorpusnye russko-francuzskie bazy dannyh glagol'nyh form i konnektorov], Linguistica e filologia, in print.
11. *Ledenev Yu. I.* (2006), Vocabulary of the dictionary of Russian grammatical words [Slovnik slovarja nepolnoznachnyh slov russkogo jazyka], Stavropol', SGU.
12. *Lyapon M. V.* (1986), Semantic structure of the complex sentence and text. A typology of intratextual relations [Smyslovaya struktura slozhnogo predlozheniya i tekst: K tipologii vnutritekstovykh otnoshenii], Moscow, Nauka.
13. *Morkovkin V. V. ed.* (1997), Dictionary of structural words of the Russian language [Slovar' strukturnykh slov russkogo yazyka], Moscow, Lazur'.
14. *Paducheva E. V.* (2013), Negative sentence in Russian [Russkoe otritsatel'noe predlozhenie], Moscow, Yazyki slavyanskoi kul'tury
15. *Peshkovskii A. M.* (2001), Russian Syntax: a scientific viewpoint [Russkij jazyk v nauchnom osveshchenii], Moscow, URSS.
16. *Priyatkina A. F.* (1977), On the difference between the conjunction and other connecting words [Ob otlichii soyuza ot drugikh svyazuyushchikh slov], Russkii yazyk v shkole, No. 4, pp. 102–106.
17. *RG–80* — Russian grammar [Russkaya grammatika], Shvedova N. Yu. (ed.), Moscow, Nauka, 1980.
18. *Rogozhnikova R. P.* (1971), Gradation conjunctions in Russian [Gradatsyonnye soyuzy v russkom yazyke], Russkii yazyk v shkole, No. 3, pp. 84–89.
19. *Rogozhnikova R. P.* (2003), Explanatory dictionary of constructions equivalent to the word [Tolkovyi slovar' sochetanii, ekvivalentnykh slovu], Moscow, Astrel'-AST.
20. *Sannikov V. Z.* (2008), Russian syntax in the formal pragmatic perspective [Russkii sintaksis v formal'no-pragmaticheskom aspekte], Moscow, Yazyki Russkoi Kul'tury, 2008.
21. *Serebryanaya F. I.* (1972), On the question of the gradation structure [K voprosu o strukture gradacionnogo rjada], Russkii jazyk v shkole, № 2, pp. 89–93.

22. *Shvedova N. Yu.* (1960), Studies of the syntax of the Russian colloquial language [Ocherki po sintaksisu russkoi razgovornoi rechi], Moscow, Academy of Sciences of the USSR.
23. *Sigal K. Ya.* (2004), Coordinating constructions in the text: a theoretic-experimental study (based on the material of simple sentences). Doct. diss. [Sochinitel'nye konstruktsii v tekste: opyt teoretiko-eksperimental'nogo issledovaniya (na materiale prostogo predlozheniya). Dokt. diss.], Moscow, Institute of Linguistics of the Russian Academy of Sciences.
24. *Tsoi A. S.* (2008), The auxiliary words as an object of the Russian lexicography [Sluzhebnye slova kak ob"ekt russkoi leksokografii], Moscou, Publishing house of Literary Institute.
25. *Ushakov D. N. ed.* (2001), Explanatory dictionary of Russian [Tolkovyi slovar' russkogo yazyka], Moscow, Veche.
26. *Zaitseva G. D.* (1998), Semantic and syntactic features of the conjunction *ne to chtoby…, no/a (not that…but)*, [Semantiko-sintaksicheskie svojstva sojuza *ne to chtoby…, no/a*], Semantics of linguistic units, Proceedings of the VI International conference, Moscow, pp. 101–103.
27. *Zav'yalov V. N.* (2009), Morphological and syntactic aspects of conjunction structure description in modern Russian. Doct. diss. [Morfologicheskie i sintaksicheskie aspekty opisaniya struktury soyuzov v sovremennom russkom yazyke. Dokt. diss.], Vladivostok, DGU.

WORD SENSE FREQUENCY OF SIMILAR POLYSEMOUS WORDS IN DIFFERENT LANGUAGES[1]

Iomdin B. L. (iomdin@ruslang.ru)
V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences; National Research University "Higher School of Economics", Moscow, Russia

Lopukhin K. A. (kostia.lopuhin@gmail.com)
Scrapinghub, Moscow, Russia

Lopukhina A. A. (nastya-merk@yandex.ru)
V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

Nosyrev G. V. (grigorij-nosyrev@yandex.ru)
Yandex, Moscow, Russia

When words have several senses, it is important to describe them properly in dictionary (a lexicographic task) and to be able to distinguish them in a given context (a computational linguistics task, WSD). Different senses normally have different frequencies in corpora. We introduced several techniques for determining sense frequency based on dictionary entries matched with data from large corpora. Information about word sense frequency is not only useful for explanatory lexicography and WSD, but it also may enrich language learning resources. Learners of a foreign language who encounter a word similar to one of their native language are often tempted to assume that the foreign word and its equivalent have the same meaning structure. Sometimes, however, this is not the case, and the most frequent sense of a word in one language may be much less frequent for its cognate. We proposed a method for detecting such cases. Having selected a set of Russian words included into the Active Dictionary of Russian which have more than two dictionary senses and have cognates in English, we estimated the frequencies for English and Russian senses using SemCor and Russian National Corpus respectively, matched the senses in each pair of words and compared their frequencies. Thus we revealed cases in which the most frequent senses and whole meaning structures are, cross-linguistically, substantially different and studied them in more detail. This technique can be applied not only to cognates, but also to pairs of words which are usually offered by the dictionaries as the translation equivalents of each other.

Key words: semantics, lexicography, polysemy, text corpora, experiments, statistical techniques, frequency, meaning frequency

---

[1] The research of Boris Iomdin, Konstantin Lopukhin and Anastasiya Lopukhina was supported by RSF (project No. 16-18-02054: Semantic, statistic and psycholinguistic analysis of lexical polysemy as a component of Russian linguistic worldview). Parsing the Dante database of English was done by Grigoriy Nosyrev. The authors would also like to thank Leonid Iomdin and Daria Shavarina as well as the anonymous reviwers for their valuable comments.

1. Introduction

Polysemy is well known to be one of the key issues both in theoretical and computational linguistics (see e.g. Pustejovsky, 1996; Apresjan, 2000; Lin and Ahrens, 2005; Agirre and Edmonds, 2007; Kwong, 2012; Hanks, 2013; Iomdin, 2014). We know many things about word senses, but very little about their frequency distribution. Indeed, it is quite difficult to obtain such information. Recently, attention has been drawn to the fact that different senses of a word normally have different frequencies in corpora (see e.g. Iomdin, Lopukhina and Nosyrev, 2014). However, there are very few resources that provide such data.

The question of the most frequent (or predominant) sense (MFS) has been discussed for the purpose of automated word sense disambiguation task (WSD), as MFS is considered to be a very powerful heuristic, which is difficult to overcome for many WSD systems (Ide and Véronis, 1998; Navigli, 2009). Various approaches for acquiring predominant senses have been applied to English. Mohammad and Hirst (2006) make use of the category information in the Macquarie Thesaurus. McCarthy et al. (2007) propose an unsupervised approach for finding the predominant senses using a distributional thesaurus and WordNet (Fellbaum 1998). Bhingardive et al. (2015) compare the embedding of a word with all of its sense embeddings (which are produced using various features of WordNet) and obtained the predominant sense with the highest similarity. For Russian, a pilot study of MFS detection was presented by Loukachevitch and Chetviorkin (2015), who used the Thesaurus of Russian Language (RuThes-lite) to determine the most frequent sense of ambiguous nouns, verbs and adjectives with the help of monosemous multiword expressions that are related to those words. Their results are comparable to the state-of-the-art in this field—the highest accuracy rate reaches 57.4%.

The overall sense distribution of a polysemous word is a question that is rarely put in focus. For English nouns, Lau et al. (2014) proposed a topic modeling-based method of estimating word sense distributions, based on Hierarchical Dirichlet Processes and on word sense induction, probabilistically mapping automatically learned topics to senses in a sense inventory. Some information about English verb pattern frequency distributions can be found in the Pattern Dictionary of English Verbs, developed by Patrick Hanks and colleagues (http://pdev.org.uk/; Hanks and Pustejovsky, 2005; Hanks, 2008). The authors emphasize that senses are associated with patterns (collocations) and not with words and that the Pattern Dictionary provides information about the relative frequency of phraseological patterns rather than dictionary senses. Cf. also Gries et al. 2010, where frequency distributions of English verbal constructions are discussed.

For Russian, a word sense frequency acquisition method and its evaluation for nouns were presented in (Lopukhina, Lopukhin and Nosyrev, in print; Lopukhina et al., 2016). In these articles we reported on our research of this issue and introduced several techniques for determining sense frequency based on dictionary entries matched with data from large corpora. This method is based on building context representations with semantic vectors (Mikolov et al., 2013) and gives robust frequency estimates with little annotated examples available from dictionaries. Supplied with examples and collocations from the Active Dictionary of Russian (Apresjan et al., 2014), the method

achieves frequency estimation error of 11–15% without any additional labeled data. It was used to obtain sense frequencies of 440 polysemous Russian nouns.

2. Meaning frequency and foreign language learning

Sense frequency distributions might be used when learning a foreign language. Language teachers can use these data for organizing senses in bilingual dictionaries, composing basic dictionary lists, preparing specific lexicon tests, etc. This information might be especially helpful for studies of cognates and other pairs of similar words in the native language of the learner and the foreign language. Languages can have many such pairs: cognates, borrowings from one of the languages into the other, international words. On the one hand, such words are easier to learn. On the other hand, learners do not always realise that even very similar polysemic words can have very different meaning structures.

In 1928, Koessler and Derocquigny coined the term "faux amis" for pairs of identical or similar cognate words with different senses, emphasizing the importance of such pairs to be identified and properly described so that translators could avoid using such words incorrectly (Koessler and Derocquigny, 1961). For English and Russian, examples such as English *artist* 'painter' vs. Russian *artist* 'performer', *English box* 'container' vs. Russian *boks* 'boxing', English *clay* 'wet soil' vs. Russian *klej* 'glue' are often given. Ample literature is devoted to such cases in different languages (see e.g. Darbelnet, 1970; Walter, 2002; Szpila, 2006; Vrbinc and Vrbinc, 2014, to name a few). In many cases, cognates in a language pair do have some senses in common, but the meaning structure is different. An important case for language learners is when the most frequent sense of a word in one language turns out to be much less frequent for its cognate.

The new techniques for counting word sense fequency on large corpora provide data for resources where similar words in different languages can be analyzed according to the comparative frequency of their senses. Bilingual comparative lexical entries sorted according to sense frequency will give learners new opportunities for better mastering the lexicon and the differences in using the words of both languages.

3. Data and Method

Our primary word sense inventory and source of materials for Russian was the Active Dictionary of Russian (ADR, see Apresjan et al., 2014). This dictionary was chosen for three reasons. First, it is the most up-to-date and most developed explanatory dictionary of Russian. Second, ADR uses a systematic approach to polysemy. The main unit of the ADR, the lexeme, is a well-established word sense identified by a set of its unique properties (syntactic, semantic and pragmatic features, sets of synonyms, analogues, antonyms and semantic derivatives etc.). Third, for each lexeme the ADR provides many sample phrases and sentences, which contributes to the precision of our word sense frequency counting technique. We selected 66 Russian nouns having phonetically similar counterparts in English. These pairs included authentic cognates of the same Indo-European

origin (such as *brat—brother*, *gus'—goose*), English borrowings into Russian (such as *bar < bar*, *biznes < business*), French borrowings into both English and Russian (such as *anekdot—anecdote*, *batareja—battery*) and words of Latin and Greek origin, mostly borrowed into Russian through French or German (such as *advokat* 'lawyer'*—advocate*, *al'bom—album*, *garmonija—harmony*). Since etymology is not relevant for our purposes, we refer to all word pairs under discussion as cognates for the sake of simplicity.

Sense frequencies of the Russian nouns were estimated automatically by performing word sense disambiguation on contexts sampled from the corpus and then calculating relative sense frequencies in the sample. For the purpose of the current study, we have sampled 1000 random contexts for each word from the domain-neutral Russian National Corpus (RNC, ruscorpora.ru, 230 million tokens in the main corpus).

The method consists of two parts: a context representation technique and a disambiguation method. We used semantic vectors as a basis for context representation and averaged them. We gave more weight to words that occur more frequently with the target word than without it. This weighting allows to capture the most important context words and build a better sense representation. Disambiguation was performed in the following way: for each sense we took all examples and collocations from the ADR and averaged their context vectors, obtaining a sense vector. During disambiguation, the context was assigned to the sense with the closest sense vector using cosine similarity measure.

Word vectors were produced using the word2vec skip-gram model with negaive sampling (Mikolov et al., 2013) from a large corpus (about 2 billion words from RuWac (Sharoff, 2006), lib.ru and Russian Wikipedia) with lemmatization, which is especially important for Russian because of its rich morphology. The dimension of word vectors was set to 300. Context vectors were calculated as a weighted average of word vectors in the fixed window of 10 words before and after the target word, where weights were proportional to PMI of context words.

The method was evaluated on hand-tagged contexts for 20 polysemous words from the ADR. It reaches an average disambiguation accuracy above 75% and a maximum frequency error below 16%. The evaluation showed that our method can perform sense frequency estimation with high accuracy (details in Lopukhina, Lopukhin and Nosyrev, in print; Lopukhin and Lopukhina 2016; Lopukhina et al., 2016).

Sense frequencies of the English counterparts were obtained from the largest sense tagged SemCor 3.0 corpus (Miller et al., 1993). SemCor is composed of 220,000 words taken from the Brown corpus (Francis and Kučera, 1979). Approximately half of the words in this corpus are open-class words (nouns, verbs, adjectives and adverbs) which have been linked to WordNet 3.0 senses (Fellbaum, 1998) by human taggers using a software interface. SemCor (and WordNet-like resources, in general) is often criticized for its excessively fine-grained sense distinction that is not supported by syntactic, syntagmatic or semantic criteria, and is neither really needed for NLP tasks (Hanks and Pustejovsky, 2005; Navigli, 2006; Snow et al., 2007) nor reflects the way people represent word meaning (Ide and Wilks, 2007; Brown, 2008). Nonetheless, SemCor remains the state-of-the-art resource in most WSD experiments. For this study, we selected a subset of words that occur at least 20 times in SemCor 3.0 and calculated the frequencies of their senses directly from labeled SemCor contexts, with an estimated maximum frequency error of 15–20%.

4. Results and Discussion

Comparing the senses of cognates by taking into account their frequency, one can detect various cases of cognates whose meaning structures are dissimilar. For some words, no senses have a match in the other language at all (these are authentic "faux amis"). In our data, typical examples of such pairs would be *arka* (most frequent sense 'a structure with a curved top and two straight sides that you can walk through, an arch') vs. *arc* 'a curved shape', *vagon* 'railway carriage' vs. *wagon* 'any of the various kinds of wheeled vehicles drawn by an animal or a tractor', *gradus* 'degree' vs. *grade* 'a level of school'. In other cases, there is one matching sense (or more), but the most frequent senses differ drastically. Cf. *avtoritet* 'the property of someone such that people see it proper to take into account his/her opinions because of his/her knowledge and experience' vs. *authority* 'the power or right to give orders or make decisions', *akcija* 'one of the equal parts of a company that you can buy, a share' vs. *action* 'something done', *artist* 'someone who performs in plays and films' vs. *artist* 'someone who makes paintings, sculptures etc', *banda* 'a group of criminals acting together' vs. *band* 'a group of musicians'.

In many other pairs, several senses match but others do not, and learners naturally tend to use them incorrectly. The Russian noun *blok* has nine senses in ADR, whereas the English noun *block* has fourteen senses in the MacMillan dictionary. Some of the senses these words share differ significantly in their frequency, e.g. 'a solid piece of something (usually having flat rectangular sides)' (more frequent in English), 'a number or quantity of related things dealt with as a unit' (more frequent in Russian), 'a pulley' (more frequent in Russian). Some other senses of the Russian word absent in its English counterpart are 'a set of tightly packed or fastened homogenous objects' (cf. English *carton*), 'a group of organizations sharing a joint purpose' (cf. English *bloc*). The senses of the English word absent in its Russian counterpart include 'a rectangular area in a city surrounded by streets and usually containing several buildings' (cf. Russian *kvartal*), 'a three-dimensional shape with six square or rectangular sides' (cf. Russian *parallelepiped*), 'a large building with a lot of different levels' (cf. Russian *mnogoaetazhka*), 'a building that is part of a larger building or group of buildings' (cf. Russian *korpus*). Another interesting example is the pair *baza* (7 senses in ADR) vs. *base* (12 senses in MacMillan, only 4 of which occur in SemCor, only 2 of the latter having correspondent senses in the Russian word).

Some of the mismatches exemplified above are not mentioned in dictionaries, and lack of knowledge thereof often leads to erroneous translations or usage. Cf. the expression *criminal authority,* clearly a calque of Russian *kriminal'nyj avtoritet*, which can be found in texts translated into English from Russian: *The Russian mass media informed that Alexander Matusov nicknamed Basmach, head of the Shelkov criminal gang, a* ***criminal authority****, has been arrested in Thailand* (Lragir.am). It occurs, however infrequently, in genuine English texts, meaning 'power to make decisions with regard to crime': *The state had no civil or* ***criminal authority*** *to force the surrender of revoked permits* (Gun control. A report by United States General Accounting Office). Another expression that reveals the Russian origin of English texts is *touristic base* meaning 'a camping site': *The* ***touristic base*** *built in Erzhei for the 'Oktai' ensemble is now open not only to the young singers but to everybody* (TuvaOnline.ru). It very rarely occurs

in genuine English texts, meaning 'a base for tourism': *Such issues need to be recognized, addressed and appropriately dealt with if a precinct is to form a truly sustainable part of a city's* ***touristic base*** (Bruce Hayllar, Tony Griffin, Deborah Edwards. City Spaces—Tourist Places); *tourist base* is slightly more frequent for this meaning.

It has to be taken into account that the usage of such words (and probably their image in the mental lexicon) may differ in bilinguals (see e.g. Schreuder and Weltens, 1993; Jiang, 2004; Dong et al., 2005, Degani and Tokowicz, 2013). By distinguishing senses that are not shared in cognate pairs in standard language, we can more easily reveal cases when they are mixed up and include them as examples of non-standard usage into standard language learning manuals or dictionaries. For example, the word *blok* is widely used by Russian immigrants in the USA in the sense of 'the distance along a city street from where one road crosses it to the next road' (absent in standard Russian, where its equivalent is the word *kvartal*): *Kogda segodnja cheloveku proshche sest' v mashinu i proexat' dva* ***bloka*** *v magazin, to emu nado pomnit' o svoem zdorov'e* 'Nowadays, when it is easier for one to drive two blocks to the store, one has to consider one's health' (Chajka, a Russian magazine published in the USA); *Tak net zhe, nado bylo emu imenno takoe mesto dlja obeda vybrat', chto by perekryt' chut' li ni samyj glavnyj vyezd iz dauntauna. Dvadcat' minut v ocheredi, chto by vyexat' iz garazha i eshche sorok, chto by proexat' dva* ***bloka*** 'But no, he had to choose precisely such a place for lunch to block probably the most imporant exit from downtown. Twenty minutes waiting in the line line to leave the garage and forty more to drive two blocks' (Livejournal.com). Many other interesting cases can be found when immigrants use a word in a sense that is absent in the standard language because it has acquired a borrowing for this sense; cf. *akcija* vs. *aekshn* < *action*: *Fil'm nudnyj, bez akcii* 'The film is boring, no action' (immigrant usage) vs. *Mogli by razbavit' specaeffektami, no i zdes' ix malo, aekshn otsutstvuet* 'They could have include special effects, but there are few of them, there is no action' (from Internet movie discussion websites).

Online dictionaries and translation memories dealing with parallel corpora sometimes contain non-genuine texts, which can result in misleading their users. E.g. the recently launched resource linguee.com, a powerful translation tool combining an editorial dictionary and a search engine for parallel corpora, provides 28 examples of parallel Russian-English texts for the Russian word *vagon*, in 5 of which it is rendered as *wagon* in English (all of them taken from Russian or Czech websites and clearly representing translations into English rather than genuine English texts), cf. *Once, when they were travelling by train, a* ***wagon*** *accidentally disconnected from the train and began to roll slowly down a slope* (from Skolkovo.ru). An inverted example: in the same dictionary we can see the English word *arc* translated into Russian as *arka* in the following sentence: *The result will be an* ***arc*** *defined by three points—V rezul'tate poluchitsja* ***arka****, postroennaja po trem tochkam* (in this geometrical context, one should use *duga* rather than *arka*). Such infelicitous translations can be found in Google Translate, too; cf. *kriminal'nyj avtoritet—criminal authority*, *sistema blokov—blocks system*, *v vagone—in the wagon* (checked on February 17, 2016).

In February 2016 we performed an online experiment with participants claiming to be (1) native speakers of English, (2) native speakers of Russian, and (3) native speakers of other languages. The respondents were to find mistakes in fifteen English

sentences taken from real texts (sometimes slightly shortened). Besides ten filler sentences (taken either from explanatory dictionaries of English or from various articles), there were five sentences containing one of the words under study (*arc*, *authority*, *base*, *block*, *wagon*):

(1) *To do so it takes a wide arc to the east, via the villages of Pilton and Croscombe.*

(2) *The tribunal had no criminal authority except over soldiers.*

(3) *The touristic base of the region expanded greatly in the 1950s.*

(4) *The band regularly played at the Mad Frog bar located just a couple of blocks from the campus.*

(5) *Families were walking beside wagons pulled by teams of oxen.*

(6) *The outhouse with an arc was built much later than the main building, in 1867.*

(7) *The head of the Shelkov gang, a criminal authority, has been arrested in Thailand.*

(8) *The touristic base built for the ensemble is now open not only to the young singers but to everybody.*

(9) *Caucasian men representing NATO block nations accused Eritrea of human rights violations.*

(10) *Once, when they were travelling, a wagon accidentally disconnected from their train.*

No further information about the nature or number of the mistakes was available. As it appeared, English speakers reported significantly less mistakes (p-value 0.03) in sentences (1–5) where the words under discussion were used in their dictionary meanings. Russian speakers reported less mistakes in sentences (6–10) where these words were used with meanings absent in English dictionaries but natural to the Russian cognates of these words. Sentences (7) and (8) were least accepted by native English speakers, many of them claimed that *criminal authority* and *touristic base* did not make sense at all. As for native Russian speakers, they had most problems in understanding sentences (2) and (3).

Table 1. Percent of mistakes in the usage of the words *arc*, *authority*, *base*, *block*, *wagon* reported by English and Russian speakers

	Sentences (1–5)	**Sentences (6–10)**
English speakers	1%	21%
Russian speakers	8%	10%

The participants of the experiment did not have to prove their proficiency in English, but for most of them it could be estimated as quite high judging by their answers where they corrected the mistakes deliberately made in the test sentences and not related to the words in question.

Hence, the results seem promising: we can obtain data that may prove useful for learners of Russian or English as well as for lexicographers and computational linguists dealing with machine translation or deep semantic analysis.

Furthermore, this technique can be applied not only to cognates, but also to pairs of words which are usually offered by dictionaries as translation equivalents of each other (see e.g. Dobrovol'skij 2007, where sets of senses of polysemous words in Russian and German are compared, for interesting observations). In order to elaborate this idea, we took a random sample of highly polysemous (more than 5 senses in ADR) and relatively frequent Russian nouns (*verx, veshch', vid, volna, vstrecha, glubina, golos*) whose most obvious English equivalents (*top, thing, view, wave, meeting, depth, voice*) are attested in SemCor, and compared their sense frequency distributions. For every word, there were senses of the Russian word that had no counterparts in the English equivalent, and vice versa. Some corresponding senses showed significant difference in frequency. Cf. *veshch'* (most common sense 'an artifact', frequency of 0.36 in Russian, cf. 0.11 for the same sense of *thing* in English) vs. *thing* (most common sense 'a special situation', frequency of 0.20, cf. 0.06 for a similar sense of *veshch'* in Russian); *vstrecha* (most common sense 'an encounter', frequency of 0.49 in Russian, cf. 0.05 for the same sense of *meeting* in English) vs. *meeting* (most common sense 'a formally arranged gathering', frequency of 0.79, cf. 0.30 for the same sense of *vstrecha* in Russian). Interestingly enough, the English word *meeting* in its most frequent sense was borrowed into Russian as *miting* in a narrower sense ('political gathering'), whereas the Russian word *vstrecha* recently developed a new sense 'an appointment'. One of the senses of *vstrecha* is 'celebration', which is apparently absent in English but can be found on Russian websites, cf. *On the chair the evenings and the celebrations devoted to the Victory Day, a meeting of New Year, to a farewell to winter (Shrovetide) etc. are spent* (Moscow State Pedagogical University, Chair of Russian as a foreign language, English website).

5. Conclusion

The idea of this study was to perform a pilot experiment determining sense frequencies for cognate Russian and English words in corpora and, taking this information into account, compare their meaning structures. The main issue is the lack of large semantically annotated corpora and dictionaries with a sufficient number of examples for each word. This limits the possibilities of automatic techniques for calculating meaning frequency. Our future plans include the following:

- applying the method of estimating word sense frequencies used for Russian on the base of the examples provided in the Active Dictionary of Russian to English, by using the data from the MacMillan dictionary and the Dante database (www.webdante.com);

- studying parallel Russian-English corpora (primarily subcorpora of the Russian National Corpus) to investigate possible differences in meaning structures of cognates used by native and non-native speakers, in Russian-to-English and English-to-Russian translations;
- creating and updating a database of Russian-English cognates comparing their senses according to their frequency;
- expanding expanding the method to other language pairs; inter alia, compare closely related languages (such as Russian and Polish) to provide material for comparative semantic and lexicological studies.

References

1. *Apresjan Ju. D.* (1974/1995). Lexical semantics. The synonymical means of language [Leksičeskaja semantika. Sinonimičeskie sredstva jazyka]. Nauka, Moscow.
2. *Apresjan Ju. D.* (ed.). (2014). Active Dictionary of Russian. A-G [Aktivnyj slovar' russkogo jazyka. A-G]. Jazyki slavjanskih kul'tur, Moscow.
3. *Agirre E. and Edmonds P.* (Eds.). (2007). Word sense disambiguation: Algorithms and applications (Vol. 33). Springer Science & Business Media.
4. *Brown S. W.* (2008). Choosing sense distinctions for WSD: Psycholinguistic evidence. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers. Association for Computational Linguistics, pages pp. 249–252.
5. *Darbelnet J.* (1970). Dictionnaires bilingues et lexicologie différentielle. Languages, 19:92–102.
6. *Degani T. and Tokowicz N.* (2013). Cross-language influences: Translation status affects intraword sense relatedness. Memory & cognition, 41(7):1046–1064.
7. *Dobrovol'skij D. O.* (2007). Polysemy structure in cross-linguistic perspectives (verbs of motion in Russian and German). In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2007", pages 162–166.
8. *Dong Y., Gui S. and MacWhinney B.* (2005). Shared and separate meanings in the bilingual mental lexicon. Bilingualism: Language and Cognition, 8(03):221–238.
9. *Fellbaum C.* (ed.) (1998). WordNet, An Electronic Lexical Database. The MIT Press, Cambridge, MA.
10. *Francis W. N. and Kučera H.* (1979). A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown). Brown University. Providence, Rhode Island.
11. *Gries S. T., Hampe B. and Schönefeld D.* (2010). Converging evidence II: More on the association of verbs and constructions. In: Empirical and experimental methods in cognitive/functional research, CSLI Publications, pages 59–72.
12. *Hanks P.* (2008). Mapping meaning onto use: a Pattern Dictionary of English Verbs. In Proceedings of the AACL, Utah.
13. *Hanks P.* (2013). Lexical analysis: Norms and exploitations. Boston: MIT Press.
14. *Hanks P. and Pustejovsky J.* (2005). A Pattern Dictionary for Natural Language Processing. Revue Française de linguistique appliquée, 10(2):63–82.

15. *Ide N. and Véronis J.* (1998). Introduction to the special issue on word sense disambiguation: the state of the art. Computational linguistics, 24(1):2–40.
16. *Ide N. and Wilks Y.* (2007). Making sense about sense. In Word sense disambiguation. Springer Netherlands: 47–73.
17. *Iomdin B.* (2014). Polysemous words in and out of the context. [Mnogoznachnyje slova v kontekste i vne konteksta]. Voprosy jazykoznanija [Issues in Linguistics]. Vol. 4. Moscow.
18. *Iomdin B., Lopukhina A., Nosyrev G.* (2014). Towards a word sense frequency dictionary. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2014", pages 205–219.
19. *Jiang N.* (2004). Semantic transfer and development in adult L2 vocabulary acquisition. In Vocabulary in a second language: Selection, acquisition, and testing, pages 101–126.
20. *Kilgarriff A., Rychly P., Smrz P. and Tugwell D.* (2004). The Sketch Engine. In Information Technology Research Institute Technical Report Series, pages 105–116.
21. *Koessler M. and Derocquigny J.* (1961). Les faux amis ou les pièges du vocabulaire anglais. Paris: Vuibert.
22. *Kusal K.* (2006). Russian-Polish interlinguistic homonymy and paronymy. A doctoral dissertation.
23. *Kwong O. Y.* (2012). New perspectives on computational and cognitive strategies for word sense disambiguation. New York: Springer Science & Business Media.
24. *Lau J. H., Cook P., McCarthy D., Gella S. and Baldwin T.* (2014). Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In Proceedings of ACL, pages 259–270.
25. *Lin C. C. and Ahrens K.* (2005) How many meanings does a word have? Meaning estimation in Chinese and English. In Language acquisition, change and emergence: Essays in evolutionary linguistics. Hong Kong, pages 437–464.
26. *Lopukhin K., Lopukhina A.* (2016). Word sense disambiguation for Russian verbs using semantic vectors and dictionary entries. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2016".
27. *Lopukhina A., Lopukhin K. and Nosyrev G.* Automated word sense frequency estimation for Russian nouns. In Quantitative Approaches to the Russian Language. In print. (Available online: sensefreq.ruslang.ru/download/Automated_Word_Sense_Frequency_Estimation_for_Russian_Nouns_Lopukhina_et_al.pdf)
28. *Lopukhina A., Lopukhin K., Iomdin B. and Nosyrev G.* (2016). The taming of the polysemy: automated word sense frequency estimation for lexicographic purposes. In Proceedings of EURALEX-2016. In print.
29. *Loukachevitch N. and Chetviorkin I.* (2015). Determining the Most Frequent Senses Using Russian Linguistic Ontology RuThes. In Proceedings of the workshop on Semantic resources and semantic annotation for Natural Language Processing and the Digital Humanities at NODALIDA.
30. *McCarthy D., Koeling R., Weeds J. and Carroll J.* (2007). Unsupervised acquisition of predominant word senses. Computational Linguistics, 33(4):553–590.
31. *Mikolov T, Chen K., Corrado G. and Dean J.* (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

32. *Miller G. A., Leacock C., Tengi R. and Bunker R. T.* (1993). A semantic concordance. In Proceedings of the workshop on Human Language Technology, Association for Computational Linguistics, pages 303–308.
33. *Mohammad S. and Hirst G.* (2006). Determining Word Sense Dominance Using a Thesaurus. In EACL.
34. *Naumov V. G.* (2014). The Ruthenian-Russian interlingual formal-semantic similarity in lexicographical representation: principles of a learner's dictionary. In: Rusin, 4(38).
35. *Navigli R.* (2006). Meaningful clustering of senses helps boost word sense disambiguation performance. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pages 105–112.
36. *Navigli R.* (2009). Word sense disambiguation: A survey. ACM Computing Surveys (CSUR) 41.2: 10.
37. *Pustejovsky J.* (1996). Lexical semantics: The problem of polysemy. Oxford.
38. *Schreuder R. and Weltens B.* (Eds.). (1993). The bilingual lexicon (Vol. 6). John Benjamins Publishing.
39. *Sharoff, S.* (2006). Creating general-purpose corpora using automated search engine queries. In Baroni and Bernardini, 2006, pp. 63–98.
40. *Snow R., Prakash S., Jurafsky D. and Ng A. Y.* (2007). Learning to merge word senses. In Proceedings of EMNLP. Prague, Czech Republic.
41. *Szpila G.* (2006). False friends in dictionaries. Bilingual false cognates lexicography in Poland. International Journal of Lexicography, 19(1):73–97.
42. *Vrbinc M. and Vrbinc A.* (2014). Friends or Foes? Phraseological False Friends in English and Slovene. AAA: Arbeiten aus Anglistik und Amerikanistik, pages 71–87.
43. *Walter H.* (2002). Les "faux amis" anglais et l'autre côté du miroir. La linguistique, 37(2):101–112.
44. *Zalizniak Anna* (2006), Polysemy in language and its representation [Mnogoznačnost' v jazyke i sposoby ee predstavlenija]. Jazyki slavjanskih kul'tur. Moscow.

ENTITY BASED SENTIMENT ANALYSIS USING SYNTAX PATTERNS AND CONVOLUTIONAL NEURAL NETWORK

Karpov I. A. (karpovilia@gmail.com)[1],
Kozhevnikov M. V. (kozhevnikov1511@gmail.com)[2],
Kazorin V. I. (zhelyazik@mail.ru)[2],
Nemov N. R. (nemo_1@pisem.net)[2]

[1]National Research University Higher School of Economics, Moscow, Russia

[2]Research and Development Institute «Kvant», Moscow, Russia

This paper provides an alternative method to extracting object-based sentiment in text messages, based on modified method previously proposed by Mingbo [8], in which we first parse the syntax, and then correlate the sentiment with the object of analysis (also referred to as entity by some, therefore, used in this article interchangeably). We show two approaches for the sentiment polarity classification: syntactic rule patterns and convolutional neural network (CNN). Even without domain specific vocabulary and sophisticated classification algorithms, rule-based approach demonstrates an average macro-F_1 based rank among the participants, whereas domain-specific vocabularies show a slightly higher macro-F_1 score, but still close to an average result. CNN approach uses syntax dependencies and linear word order to obtain more extensive information about object relations. Convolution patterns, designed in this approach, are very similar to rules, obtained with rule-based approach. In our proposed approach, the neural network was trained with different Word2Vec (WV) models; we compared their performance relative to each other. In this paper, we show that learning a domain-specific WV offers slight progress in performance. Resulting macro-F_1 score show performance in the into top three of the overall results among the competitors, participating in 2016 SentiRuEval event. Originally, we have not submitted our results to this competition at the time it was held, but had a chance to compare them post-hoc. We also combine the CNN approach with the rule-based approach and discuss the obtained differences in results. All training sets, evaluation metrics and experiments are used according to SentiRuEval 2016.

Keywords: sentiment analysis, object-oriented sentiment analysis, syntax patterns, machine learning, convolution neural network

ОБЪЕКТНО-ОРИЕНТИРОВАННЫЙ АНАЛИЗ ТОНАЛЬНОСТИ ПРИ ПОМОЩИ СИНТАКСИЧЕСКИХ ШАБЛОНОВ И СВЕРТОЧНОЙ НЕЙРОННОЙ СЕТИ

Карпов И. А. (karpovilia@gmail.com)[1],
Кожевников М. В. (kozhevnikov1511@gmail.com)[2],
Казорин В. И. (zhelyazik@mail.ru)[2],
Немов Н. Р. (nemo_1@pisem.net)[2]

[1]НИУ ВШЭ, Москва, Россия
[2]НИИ «Квант», Москва, Россия

Ключевые слова: определение тональности, тональность объектов, синтаксические шаблоны, машинное обучение, сверточные нейронные сети

1. Introduction

The online reputation analysis task, performed on social networks' data such as Twitter data, has several differences from the traditional sentiment tasks. We suggest that performance of systems designed to solve this problem depends on three factors:

(i) **Lexicon actualization**—the first issue is that there are many texts that do not contain any intuitively subjective words, but nonetheless, express a person's attitude. Usually such words are domain-specific. For example, in the context of everyday media usage of the word, the verb "*выдавать*" (most closely translated as to "fib"), has negative sentiment, because it is frequently used in meaning "*to lie*" ("*представлять что-либо не тем, чем оно является на самом деле*") or "*to betray*" ("делать донос, предавать")[1]. However, in banking, the same word means "*to issue a credit card*" or "*provide a loan*" ("*передавать в чье-л. распоряжение*") and usually has a positive sentiment. A promising approach to sentiment word extraction was described in [1]. In this paper, we study different Word2Vec models, trained by using news and social networks data.

The second issue is that pejorative lexicon used by social network users does not always indicates a negative opinion. For example, someone may be using swear language to indicate either negative or positive affect, which may not be obvious immediately.

[1] Meanings of the verb "*выдавать*" are provided by WikiDictionary: http://ru.wiktionary.org/wiki/выдавать

(ii) Object matching—the issue here is linking the sentiment word with key object, especially when text is long or when there are multiple entities mentioned. For example, it would be very difficult to analyze the sentence "***Билайн**, которым я пользовался два года, гораздо **лучше МТС***" ("*Beeline, that I've used for two years, is much better than MTS*") using only linear context because key sentiment word "*лучше*" is much closer in absolute word distance to the object "*MTS,*" rather than "*Beeline*." Also, according to our experience, analysis of comparison structures such as "*A is better than B,*" without syntax information, produces erroneous results. Both our approaches incorporate syntax information, as described later in the paper. With that, we are basing our method on the classical approaches to solving this problem as described in [9], [11].

(iii) Subjective fact interpretation—recent sentiment evaluation competitions show tendency of adding fact interpretations to sentiment analysis. For example, in the sentence "***Сбербанк** подаст в суд иск по банкротству **Мечела***" ("***Sberbank** will bring a bankruptcy case against **Mechel** to court*"), we have a fact of a bankruptcy, negative for "*Mechel*", but an ordinary bank activity for "*Sberbank*." Processing such data requires many specific, often counteractive rules to deal with the problem of contradicting sentiments in the traditional rule-based approach, but could be efficiently performed by modern neural networks.

Recent works involving CNN-based approaches in English [8], [4], [2] have demonstrated excellent results on various classification tasks, including sentiment analysis. Because we expected that (ii) and (iii) factors could only be solved with syntax-dependency information, we used CNN, which uses not only linear word order, but also syntax dependencies to extract sentiment, and could allow for more efficiency in the task processing.

Rule-based approach, described later in this paper, is similar to the RCO approach [4], but there are differences in text preprocessing and lexical dictionaries' extraction. The CNN approach is also similar to [8] paper, but we have changed the input vector and made entity token with special TARGET mark to achieve a more efficient object-oriented sentiment analysis. We also used custom convolution patterns in this work.

2. Methods

Figure 1 gives a brief overview of the proposed approach. Input text is parsed with graphematic, morphological, and syntax parsers at the Text preprocessing stage. The rule-based approach assumes that predefined syntax patterns are enhanced by preliminarily generated Word2Vec models and sentiment dictionaries. as Resulting feature vector is analyzed by the extremely naive classifier that labels the object sentiment according to quantity of sentiment facts, linked with this object in the text. The resulting sentiment is a net sum of positive and negative sentiment labels. In case of the CNN-based approach, preprocessed text is vectorized with preliminary generated Word2Vec model. CNN returns the sentiment label as a result. We first build two separate classifiers, which can be easily combined, as shown in experiments section later in the paper. We now discuss each module in detail.

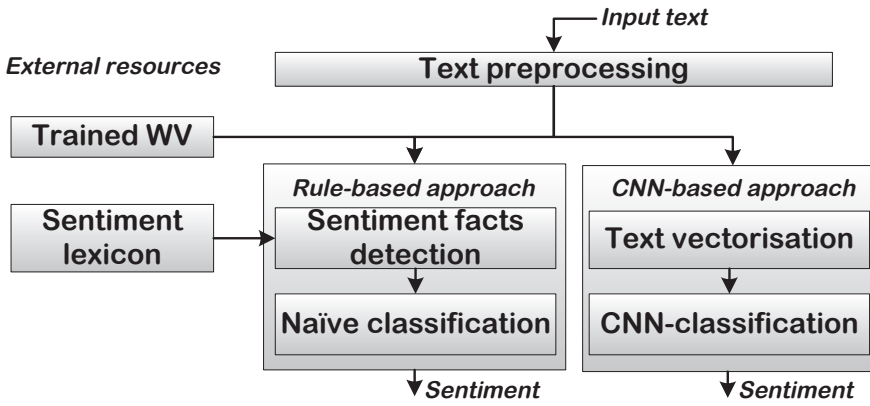


Fig. 1. Overall system architecture

2.1. Text preprocessing

Since input data from social networks is very noisy, a substantial amount of preprocessing is required. These steps are discussed below.

2.1.1. Remove URLs

URLs do not carry a lot of substantial information regarding the sentiment of the tweet and contaminate the dictionary, so we remove them with simple regex.

2.1.2. Remove nontextual data

Hashtags and tokens, starting with an "at" sign (@) represent important information about the reviewed object. In order to find it, we remove certain punctuation such as quotation marks, hyphens, asterisks, "at" signs, etc.

2.1.3. Tokenisation & morphology

We applied our own NLP toolkit [3] and Mystem parser developed by Yandex[2] for text preprocessing. Morphological analysis shows similar results, but tokenization, done by Mystem, was not designed to handle emotions and other punctuation specifics of social networks, so we preferred our own parser, which could overcome these limitations.

2.1.4. Named Entity (NE) recognition

We used Wikipedia hyperlink structure to find entities and their possible occurrences in the text as proposed in [12]. The basic algorithm was enhanced by adding transcripts and translations for each separately occurring appearances of key objects. We also generate separate grammatical cases for each normal form of the word

[2] https://tech.yandex.ru/mystem/

or phrase, describing the key object, and add them as a possible occurrence of key object in the text. As a result, we formulate the dictionary of the key objects' occurrences in the text. During the text processing step we replace key objects' occurrence with a special "TARGET" token and an appropriate morphology information.

2.1.5. Syntax parsing

We process entire dataset with malt parser [10], trained on our own news corpora to get dependency trees used by both approaches. If the tweet contains multiple root nodes, they are all added as descendants of special fake "*ROOT*" node. Sample syntax parse result is shown at figure 2. In general, our constructs had a single root, but in case it was not so, we used the described approach.

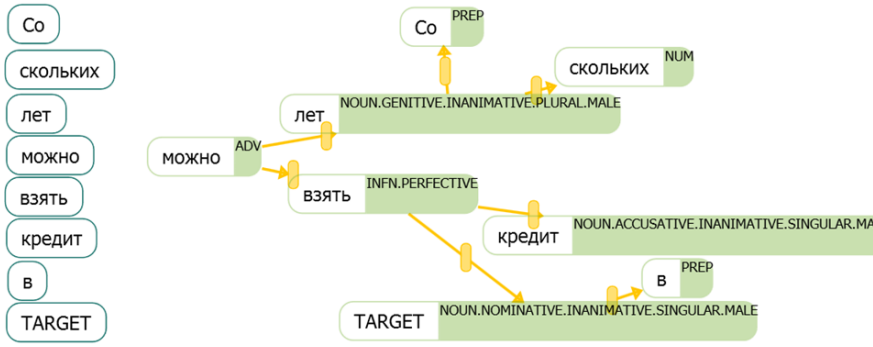


Fig. 2. Syntax parse result example

2.2. Word2Vec training

We use the Word2Vec (VW) [7] modeling in both the rule- and the CNN- based approaches. In case of a rule-based approach, WV is used for computing semantic similarity between sentiment words. In CNN, WV is needed to represent text as a matrix for the neural network input. WV is trained on word lemmas with part-of-speech codes. We exclude punctuation, conjunctions, prepositions, particles and short (less than 3 symbols) English words from the training data. We use 300-dimension vector size skip-gram model with the minimum cut-off for the number of words = 3 in all cases.

Corpora lexicon plays an important role in generating WV model. We gathered nearly 1.5 million twitter search results about general topics such as music, cinema, travelling, literature, sports, etc[3]. Obtained model takes into account the specifics of twitter language, but still suffers from the word sense ambiguity problem. Therefore, we also gathered twitter search results for banking and telecom topics of nearly 100,000 tweets each.

[3] Selected categories list, trained models and project code can be found at http://github.com/lab533/RuSentiEval2016

The following combinations of gathered corpora were made to find the balance between corpora size and word ambiguity problem:

- WV_Banks_clear: 120,000 bank tweets
- WV_TTK_ clear: 120,000 telecom tweets
- WV_Twitter: 1,500,000 gathered twits
- WV_news: 4,500,000 news texts

We also added news-based WV to explore the role of twitter-specific vocabulary in sentiment tasks. Different mixtures of gathered corpora was evaluated as described in experiments section.

2.3. Rule-based approach

As a first step, we look for sentiment words of a tweet. We use our own universal dictionary of sentiment words for this purpose. Dictionary consists of 2,074 positive and 6136 negative normal word forms, manually verified by experts. After inflection of normal words forms and their enrichment with top 2 most similar WV words, dictionary was transformed to 60,288 positive and 189,953 negative word forms. Using the syntax tree of the sentence, which contains sentiment word, we detect modal verbs and negotiation markers (like "*не*", "*нет*" etc.).

Next, we define sentiment facts associated with sentiment words. Sentiment fact is a semantically isolated part of a syntactic tree, which contains the sentiment word. In our rule-based approach, there are two types of sentiment facts, depending on parent of the sentiment word. If a parent of the sentiment word is a subordinate part of a sentence, a sentiment fact is a branch of the syntax tree with the parent of the sentiment word. This is the first type of sentiment facts. An example of such fact is the phrase "*уродливое здание Сбербанка*" ("*ugly Sberbank building*") of the sentence "*В каждом городе России есть уродливое здание Сбербанка*" ("*There is an ugly Sberbank building in each city in Russia*"), as shown at figure 3.

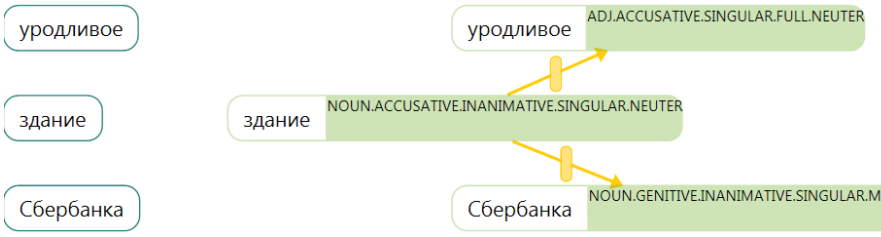


Fig. 3. Sentiment fact with an adjective modifier sentiment word

The second type of a sentiment fact is the sentiment word or its parent, which is one of the subjects of the sentence or one of its predicates. In this case, the sentiment fact includes a predicate, a subject, and all of their children tokens. For example,

the sentiment fact here is the "*ненавижу Райффайзен банк*" ("*hate Raiffaizen bank*") in the sentence "*Я не устану повторять, что ненавижу Райффайзен банк*" ("*I will never stop saying that I hate Raiffaizen bank*"), as shown at figure 4.

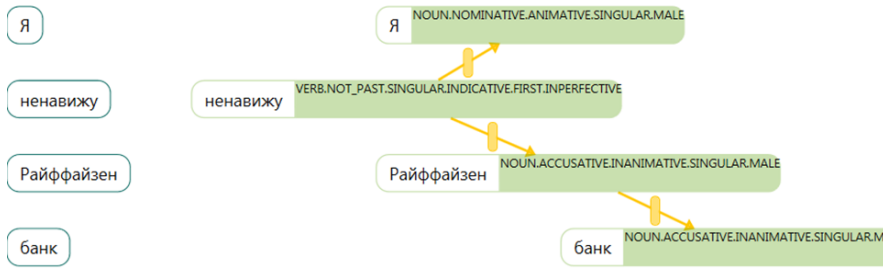


Fig. 4. Sentiment fact with a predicate sentiment word

Next, we unite neighboring sentiment facts: if one of the tokens of the sentiment fact has a syntactic connection with a token of another fact, these two facts get combined into one. Then we apply rules of combination of positive and negative sentiment words inside facts, and calculate integer sentiment index for each fact.

To improve general performance of the algorithm, we also made some individual rules for each domain:

- Stop-words list (words from dictionary that do not have any sentiment for a specific domain);
- Unigram and n-gram words list (words that have a sentiment value only for a specific domain);
- Applying "No-rule" (words or n-grams that have sentiment only with or without negotiation);

Finally, we find sentiment facts that contain a target object. If there is no sentiment fact with a target object, we assign object to the nearest fact in the syntactic tree. Then we calculate total sentiment score for each object and use it as a final sentiment result. We mark tweets that do not have any sentiment facts as neutral.

2.4. Convolutional neural network approach

Convolutional neural networks (CNNs), originally invented in computer vision [5], in recent years have been applied in many natural language processing (NLP) tasks such as authorship detection, question answering, and sentiment analysis. Let $x_i \in R^k$ be the k-dimensional word vector corresponding to the i-th word in the sentence. The sentence of length n can be described as

$$x_{1:n} = x_1 \oplus x_2 \oplus \ldots \oplus x_n \quad (1)$$

where $\oplus$ is the concatenation operator. Such vector is considered to be CNN input. A convolution operation involves a filter $w \in R^{hk}$, which is applied to a window of h words to produce a new feature. This filter is applied to each possible window of words in the sentence to produce a feature map. Max-overtime pooling [4] operation over the feature map is applied to capture the most important feature—one with the highest value—for each feature map. These features maps form the penultimate layer and are passed to a fully connected softmax layer, whose output is the probability distribution over labels.

2.4.1. Dependency-based Convolution

We are using the Mingbo's [8] approach to include syntax information into the classification process, where dependency-based convolution is described as follows:

$$x_{1:n} = x_i \oplus x_{p(i)} \oplus \ldots \oplus x_{p^{n-1}(i)} \tag{2}$$

where $p^n(i)$ returns the i-th word's n-th parent, which is recursively defined as:

$$p^n(i) = \begin{cases} p\left(p^{n-1}(i)\right) & if\ n > 0 \\ i & if\ n = 0 \end{cases} \tag{3}$$

Text preprocessing notation and the peculiarities of twitter text often cause the TARGET node to be separated from the sentiment fact into a different sentence. In order to capture these long-distance dependencies in the entire tweet, we use sibling convolutions defined as:

$$s(i,j) = \begin{cases} 1 & if\ p(i) = p(j) \\ 0 & if\ p(i) \neq p(j) \end{cases} \tag{4}$$

where $i > j$. We take maximum five first left siblings of i-th token to avoid combinatorial explosion.

2.4.2. Convolution patterns

Inspired by rule-based approach, we added several convolution patterns of length two to four words. Maximum pattern length was taken from the rule-based approach, where we have very few patterns longer than four tokens deep. It should be mentioned that one token doesn't equal one word, because we replace phrases with TARGET mark during object matching phase.

Table 1. Tree convolution patterns of different depth

Pattern depth	Pattern
2	
2	

Pattern depth	Pattern
3	
3	
3	
4	
4	

Asterisk in table 1 means that information about this word is not included to a convolution pattern. We also add information about the sequential token order in the tweet to compensate for parsing errors during the syntax analysis stage. The final input vector is a concatenation of feature maps from tree-based information and n-grams, with n=5.

2.4.3. Training

We substitute all "word + POS" pairs are by unique ids and align all sentences to length 50 (zero padding). We take first 5 anchestors and first 5 siblings for each word in a sentence and concatenate all words to form input vector for our NN. Neural network consists of the following layers:

- embedding layer—to turn word ids to word vectores, we used only words, contained in training;
- convolution layer—layer with rectified linear unit (ReLU) activation where convolution patterns are applied as described in table 1;
- maxPooling layer—which is down-sampling convolution layer output;
- dropout layer—with dropout rate was set to 0.25;
- dense layer—with ReLU activation;
- dropout layer—with dropout rate was set to 0.5;
- softmax layer—to form classification output.

We employ random dropout on penultimate layer to avoid overtraining as described in [4]. We trained our CNN for 40 epochs, but did not observe any increase in quality after the 2th epoch. Training was done through stochastic gradient descent over shuffled mini-batches with the AdaGrad update rule. Trained CNN models with exact parameters could be found at project repository, noted at section 2.2.

3. Experiments

Results of our evaluation are presented in Table 2. Consistent with standards of the RusSentiEval, the macro-averaged F_1-measurewas used as a primary evaluation metric [6]. Table 2 below describes positive and negative sentiment classes and micro-averaged F_1.

Table 2. Performance of rule- and CNN-based approaches in different configuration

Domain	Approach	Training collection	WV	F_1 positive	F_1 negative	Macro-average F_1	Micro-average F_1
Banks	Rule-based	Banks	—	0.387	0.501	0.443	0.463
	Rule-based with domain rules	Banks	—	0.394	0.524	0.459	0.482
	CNN	Banks	Random	0.425	0.555	0.490	0.523
		Banks	News	0.422	0.555	0.489	0.523
		Banks	Twitter	0.429	0.552	0.490	0.522
		Banks & TTK	Random	0.446	0.618	0.532	0.574
		Banks & TTK	News	0.455	0.611	0.533	0.572
		Banks & TTK	Twitter	0.456	0.615	**0.536**	0.574
Telecom	Rule-based	TTK	—	0.280	0.682	0.481	0.569
	Rule-based with domain rules	TTK	—	0.285	0.695	0.490	0.582
	CNN	TTK	Random	0.097	0.556	0.326	0.497
		TTK	News	0.091	0.557	0.324	0.499
		TTK	Twitter	0.091	0.559	0.325	0.500
		Banks & TTK	Random	0.307	0.738	0.523	0.681
		Banks & TTK	News	0.298	0.740	0.519	0.682
		Banks & TTK	Twitter	0.313	0.739	**0.526**	0.682

In the table above, the column “Training collection” describes the collection, chosen to train the model. In case of “Banks & TTK” value, model was trained on both Banks and Telecom data shuffled in random order. “WV” column describes Word2Vec model, used in the experiment. Results in Table 2 demonstrate that training corpora size is more important than the selected VW model. It also appears that WV is extremely sensitive to the input data. In our case VW, trained with only the domain specific data, shows better results that can be increased by acquiring bigger corpora.

3.1. Overall Performance

The evaluation metric used in the SentiRuEval 2016 competition is the macro-averaged F_1 measure calculated over the positive and negative classes. Table **3** shows the overall performance of our system for bank and telecom datasets.

Table 3. Performance of our method and best F_1 measure among all participants

Domain	Approach	F_1 positive	F_1 negative	Macro-average F_1	Micro-average F_1
Banks	Rule-based	0.394	0.524	0.459	0.482
	CNN	0.456	0.615	0.536	0.574
	Hybrid	0.457	0.619	**0.538**	**0.577**
	SentiRuEval best			0.552	
Telecom	Rule-based	0.285	0.695	0.490	0.582
	CNN	0.313	0.739	0.526	0.682
	Hybrid	0.313	0.740	**0.527**	**0.684**
	SentiRuEval best			**0.559**	

In case of rule-based approach, the system was not developed for banks or telecom companies' domains specially. Rule-based approach did not use any machine learning. Training collection was used only for extracting the proposed domain-specific rules, which approximately increased macro-average F-measure by 0.015.

With the Hybrid approach, final sentiment marks of neutral tweets, gained from rule-based approach, are inputs for a CNN. In general, rules give more precise result, but fail in recall. This method shows small performance progress in case of telecom domain, but does not help in bank domain, which may be caused by overfitting when multiple rules interfere each other.

4. Conclusions

We presented results of sentiment analysis on Twitter by building two approaches based on hand-written syntactic rules and CNN. Rule-based linguistic method showed average performance result, which makes it useful when training collection is not available. Few hand-written rules with well-filtered dictionaries can give a little boost to the CNN output, but the system degrades as rules count increases. CNN show very high quality result that coincides with the best results of the competition, but this approach requires relatively large training collections. The same problem occurs in distributive semantics, applied in this work. Word2vec can extract deep semantic features between words if training corpora is large enough.

Acknowledgment

The article was prepared within the framework of a subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the Global Competitiveness Program.

References

1. *Chetviorkin I., Loukachevitch N.* (2012), Extraction of Russian Sentiment Lexicon for Product Meta-Domain. Proceedings of the 26nd International Conference on Computational Linguistics (Colling-2012), Mumbai, pp. 593–610.
2. *Kalchbrenner N., Grefenstette E., Blunsom P.* (2014), A Convolutional Neural Network for Modelling Sentences, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL-2014), Baltimore, pp. 655–665.
3. *Karpov I., Goroslavskiy A.* (2012) Application of BIRCH to text clustering. Proceedings of the 14th All-Russian Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" (RCDL-2012), Pereslavl Zalesskii, pp. 102–105.
4. *Kim Y.* (2014), Convolutional neural networks for sentence classification, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, pp. 1746–1751.
5. *LeCun Y., Cortes C., Bottou L., Jackel L.* (1995), Comparison of Learning Algorithms for Handwriting Digit Recognition, International Conference on Artificial Neural Networks, Paris, pp. 53–60.
6. *Loukachevitch, N. V., Blinov, P. D., Kotelnikov, E. V., Rubtsova, Y. V., Ivanov, V. V., Tutubalina, E.* (2015), Sentirueval: Testing Object-Oriented Sentiment Analysis Systems in Russian, Proceedings of the International Conference "Dialog 2015", Moscow.
7. *Mikolov T., Chen K., Corrado G., Dean J.* (2013), Distributed Representations of Words and Phrases and their Compositionality, Proceedings of Advances in Neural Information Processing Systems 26 (NIPS 2013), Tahoe, pp. 3111–3119
8. *Mingbo M., Liang H., Bing X., Bowen Z.* (2015), Dependency-based Convolutional Neural Networks for Sentence Embedding, Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics (ACL-2015), pp. 174–179.
9. *Nivre J., Iomdin L. L., Boguslavsky I. M.* (2008), Parsing the SynTagRus Treebank of Russian, Proceedings of the 22nd International Conference on Computational Linguistics (Colling-2008), Manchester, pp. 641–648.
10. *Nivre, J., Hall, J., Nilsson J., Chanev A., Eryigit G., Kübler S., Marinov S.* (2005), MaltParser: A language-independent system for data-driven dependency parsing, Natural Language Engineering, Vol. 13, № 1 pp. 95–135
11. *Polyakov P. Y., Kalinina M. V., Pleshko V. V.* (2015), Automatic Object-Oriented Sentiment Analysis by Means of Semantic Templates and Sentiment Lexicon Dictionaries, Proceedings of the International Conference "Dialog 2015", Moscow.
12. *Sachidanandan S., Sambaturu P., Karlapalem K.* (2013), NERTUW: Named entity recognition on tweets using Wikipedia, Concept Extraction Challenge Proceedings, Rio de Janeiro, pp. 67–70.

LARGE CORPORA AND FREQUENCY NOUNS

Khokhlova M. V. (m.khokhlova@spbu.ru)

St. Petersburg State University, St. Petersburg, Russia

The paper describes a new branch in corpus linguistics that deals with building and using large corpora. We introduce several new large Russian corpora that have recently become available. The paper gives a survey of the given corpora and analyzes a number of Russian nouns across the following corpora of different sizes: the Russian Web corpus by S. Sharoff (187.97 mln tokens), ruTenTen (18.28 bln tokens) and its sample (1.25 bln tokens). The research focuses on the discussion on these corpora, their comparison and the study of frequency properties for the high- and low frequency Russian nouns comparing them with data published in the Frequency Dictionary. The analysis shows the lists presented in the frequency dictionary of Russian differs from the corpus data depending on types of the nouns.

Key words: text corpora, web corpus, frequency, frequency nouns, statistics, big data

БОЛЬШИЕ КОРПУСА И ЧАСТОТНЫЕ СУЩЕСТВИТЕЛЬНЫЕ

Хохлова М. В. (m.khokhlova@spbu.ru)

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Статья посвящена использованию больших корпусов, которые стали активно развиваться в последнее время. В ней представлены результаты исследования частотных существительных русского языка на материале корпусов разных объемов. В статье дается обзор русских корпусов большого объема. Обсуждаются различия между корпусами и характеристики высоко- и низкочастотных существительных, проводится их анализ с данными частотного словаря русского языка. Анализ показывает, что данные, приведенные в словаре, и результаты, полученные на корпусной основе, отличаются для разного типа существительных.

Ключевые слова: корпус текстов, Интернет-корпус, частота, частотные существительные, статистика, большие данные

Introduction

The idea of corpora that contain big data has attracted scholars' attention for a long time. However, it was the availability of wide technical opportunities that gave impetus to the development of a new research field which enables researchers to collect data automatically from the Internet (see, for example [Kehoe, Renouf 2002; Kilgarriff, Grefenstette 2003; Belikov, Selegey, Sharoff 2012]). Researchers found it attractive to make statistical inferences and to verify the results on increasingly larger scope of data. 1 mln or 100 mln tokens are not thresholds for the corpora. At the same time access to big corpora provokes new challenges: what can we see with big data and how does it affect the results? Do linguists actually need large corpora or their appetites can be satisfied with less data? Can small corpora be viewed as little big ones [Sinclair 2004]?

1. Large Russian Corpora

Large corpora with volumes exceeding 100 mln tokens have appeared just recently. This idea is closely related to the technical resources and thus the gradually changing paradigm in corpus linguistics moving forward from "manual" approach to more automatic one. Nowadays one can speak about two types of corpora, some authors distinguish between three types [Belikov, Selegey, Sharoff 2012]. For the Russian language the most famous and popular corpus of the first type is the National Russian Corpus; altogether its subcorpora comprise 600 mln words. This corpus was built according to the "classic" style, i.e. linguists selected relevant texts, annotated them and included into the database. Corpora of the second type are collected automatically from the Web (frankly speaking, to a certain degree that holds true for the first type also). For the Russian language we can name "The General Internet-Corpus of Russian" that now exceeds 15,000 mln words and its authors aim at 50,000 mln words [Piperski et al. 2013]. Also, there is a number of corpora made within the Aranea project, which includes a Russian corpus called Araneum Russicum Maius & Minus [Benko 2014]. The Russian Web corpus was collected by S. Sharoff [Sharoff 2006] using the methodology for crawling web-based texts outlined in [Baroni, Bernardini 2004]. The developers use a list of 500 most frequent words in a language that are not function words and don't belong to a specific domain. At the next stage a program produces a list of queries (from 5,000 up to 8,000); each of them consists of 4 words from the list. As an output there is a new list of web links, 10 top results are used for downloading files. The next step requires further postprocessing, such as encoding correction, removing duplicates (filtering pages in other languages using the same alphabet, e.g. deleting texts written in Cyrillic in Slavic languages other than Russian) and technical information. The TenTen family [Jakubíček, Kilgarriff, Kovář, Rychlý, Suchomel 2013] includes corpora of various languages of the order of 10 billion words. The ruTenTen Russian corpus is one of the biggest among them along with the English, German, French and Spanish collections. Building these corpora implies that special attention is paid to the process of de-duplication in order to delete multiple

copies of the same chunks of texts and thus to reduce postprocessing. Also the crawler downloads texts that have only full sentences (not navigation data).

The above-mentioned technologies can be found attractive enabling the researchers to create corpora of different languages and not demanding the time-consuming stage of collecting texts (however, this advantage can be questioned if we remember about such inherent properties of a corpus and a sample as their representativeness and balance).

To our best knowledge, there are no large corpora studies of linguistic phenomena on the Russian data, which would come up with a comparative analysis of these corpora (e.g. "big" vs. "little" corpora or "manual" vs. "automatic"). A survey of the Aranea Russian corpora was made in [Zakharov 2015]. For the English corpora [Kilgarriff 2001] suggests the χ^2-test as a suitable measure for comparing corpora that outperforms other methods. For Chinese corpora an attempt of evaluation is described in [Shu-Kai Hsieh 2014]. Comparative analysis of frequency lists can be viewed as a similar issue to a certain degree. For Russian the method was proposed in [Shaikevich 2015] describing the *Cxy* measure.

2. Experiments

The aim of our research is to compare frequency properties of a number of Russian nouns across corpora of different sizes, to identify differences, and to analyze them. We selected several corpora that had been collected and built automatically—the Russian Web corpus by S. Sharoff (187.97 mln tokens), ruTenTen (18.28 bln tokens) and its sample (1.25 bln tokens). The latter is a subcorpus of the ruTenTen corpus that was randomly generated, so it comprises the same texts but differs in size from ruTenTen. To succeed in our study we studied properties of high- and low frequency nouns that had been selected from the dictionary.

2.1. High-Frequency Nouns: Selection

The majority of Russian texts in web corpora come from news websites, blogs, commercial websites, social media groups etc. Fiction texts are less common for such corpora; therefore, we decided to focus on high-frequency vocabulary that is associated with the above-mentioned functional styles. To this end, we compiled two word lists. One word list (see Tables 1 and 2) contained nouns that the Frequency Dictionary [Lyashevskaya. Sharoff 2009] ranked top by frequency in social and political journalism and non-fiction texts (the Frequency Dictionary provides separate frequency lists for both types of texts). Hence both lists include top 10 nouns for the given style.

Table 1. High-frequency nouns in non-fiction texts

No.	Lemma	Translation	Frequency (ipm)
1	god	year	4,624.2
2	vremja	time	2,080.5
3	čelovek	man	1,945.3
4	sistema	system	1,798.0
5	rabota	job	1,766.4
6	stat'ja	article	1,363.0
7	delo	affair	1,339.5
8	slučaj	case	1,259.0
9	process	process	1,221.8
10	vopros	question	1,180.9

Table 2. High-frequency nouns in texts belonging to social and political journalism

No.	Lemma	Translation	Frequency (ipm)
1	god	year	5589.5
2	čelovek	man	2950.1
3	vremja	time	2364.6
4	žizn'	life	1548.4
5	delo	affair	1482.0
6	den'	day	1397.8
7	rabota	job	1272.4
8	strana	country	1203.9
9	vopros	question	992.0
10	slovo	word	989.7

Tables 1 and 2 show that both lists overlap; therefore, we ended up with just 14 words on the final list (indexes indicate that a word is rated top by frequency for both journalism and non-fiction texts): *god* 'year'[1.2], *vremja* 'time'[1.2], *čelovek* 'man'[1.2], *sistema* 'system', *rabota* 'job'[1.2], *stat'ja* 'article', *delo* 'affair'[1.2], *slučaj* 'case', *process* 'process', *vopros* 'question'[1.2], *žizn'* 'life', *den'* 'day', *strana* 'country' and *slovo* 'word'.

The other list contains nouns that belong to the so-called style-specific vocabulary (i.e. typical) [Lyashevskaya. Sharoff 2009] for either social and political journalism or non-fiction texts (see Tables 3 and 4).

Table 3. High-frequency style-specific nouns on the word list for non-fiction texts (social and political journalism excluded)

No.	Lemma	Translation	Frequency (ipm)		LL-score[1]
			Corpus	Subcorpus	
1	stat'ja	article	395.0	1,363.0	10,512
2	sistema	system	617.8	1,798.0	9,943
3	federacija	federation	258.9	1,003.1	9,329
4	process	process	371.7	1,221.8	8,639
5	risunok	picture	179.2	776.2	8,451
6	virus	virus	106.5	584.1	8,388
7	issledovanie	study	200.5	799.6	7,762
8	ispol'zovanie	usage	190.3	757.9	7,342
9	sud	court	371.1	1,153.9	7,334
10	metod	method	197.0	772.3	7,312

Table 4. High-frequency style-specific nouns on the word list for social and political journalism

No.	Lemma	Translation	Frequency (ipm)		LL-score
			Corpus	Subcorpus	
1	prezident	president	311.0	634.6	2,186
2	teatr	theatre	305.3	611.9	1,944
3	god	year	3,727.5	5,589.5	1,435
4	spektakl'	play	164.7	350.0	1,429
5	pravitel'stvo	government	277.7	531.2	1,341
6	kompanija	company	392.7	699.0	1,149
7	strana	country	725.7	1,203.9	1,135
8	fil'm	film	196.8	380.1	1,009
9	reforma	reform	133.1	273.0	963
10	vybory	elections	117.7	243.4	889

2.2. High-Frequency Nouns: Results

In our research we have also analyzed 10 top-frequency nouns in the three corpora. The Russian Web Corpus list was almost identical to the list in [Lyashevskaya. Sharoff 2009]. The results for the two other corpora are more exciting. For example,

[1] The logarithmic likelihood score is a static measure used by the authors of the dictionary to identify style-specific vocabulary. In Tables 3 and 4 the results are arranged according to this parameter.

in ruTenTen *god* 'year', *rabota* 'job', *vremja* 'time', *čelovek* 'man', *kompanija* 'company', *sistema* 'system', *sajt* 'site', *den'* 'day', *mesto* 'place' and *Rossija* 'Russia' topped the frequency ranking. The lemmata *sistema* and *kompanija* were ranked 26 and 59 respectively on the high-frequency nouns list, whereas *sajt* and *Rossija* were entirely missing on this list. In the ruTenTen subset the word *sajt* was missing, whereas the lexeme *rebënok* 'baby', ranked 22 in the Frequency Dictionary, was present on the subset.

We referred to the three corpora to study frequencies of the words on the lists (see Tables 1 and 2); you can find the results on Table 5 and Fig. 1. as well as on Table 6 and Fig. 2.

Table 5. Frequencies of nouns on the non-fiction word list (journalism excluded) calculated as per three corpora

			Frequency (ipm)			
			Frequency word list for non-fiction (journalism excluded) in the Frequency Dictionary	Russian Web Corpus	ruTenTen	
No.	Lemma	Translation			Corpus	Sample
1	god	year	4,624.2	2,220.74	3,078.97	3,076.99
2	vremja	time	2,080.5	1,761.06	1,790.84	1,793.41
3	čelovek	man	1,945.3	2,343.79	1,955.40	1,950.79
4	sistema	system	1798	527.61	998.41	1,006.66
5	rabota	job	1,766.4	885.02	1,509.37	1,510.41
6	stat'ja	article	1363	257.55	293.72	292.09
7	delo	affair	1,339.5	1,037.09	813.12	809.29
8	slučaj	case	1259	632.16	750.61	752.11
9	process	process	1,221.8	294.37	473.94	478.05
10	vopros	question	1,180.9	853.94	866.03	869.27

Table 5 and Fig. 1 show the data for nouns in Table 1. We can see that both ruTenTen charts for the corpus and the subset are identical, which means that these words have identical distribution. The frequencies, indicated in the Dictionary, are the highest, except the frequency of the lemma *čelovek* which has the highest frequency in Russian Web Corpus. All the three corpora rank the words somewhat differently from the ranking in the Dictionary—two nouns in Russian Web Corpus have the same ranking as in the Dictionary, while ruTenTen (and the subset) contains four such nouns. Moreover, both corpora agree on the ranking of the four words *vremja*, *vopros*, *process* and *stat'ja*. Spearman's rank correlation coefficient between the ranked word lists in the Frequency Dictionary and in Russian Web Corpus is 0.61, whereas the coefficient between the Frequency Dictionary and the ruTenTen corpus stands at 0.78, which in the latter case reveals that the dictionary and the corpus have much more in common.

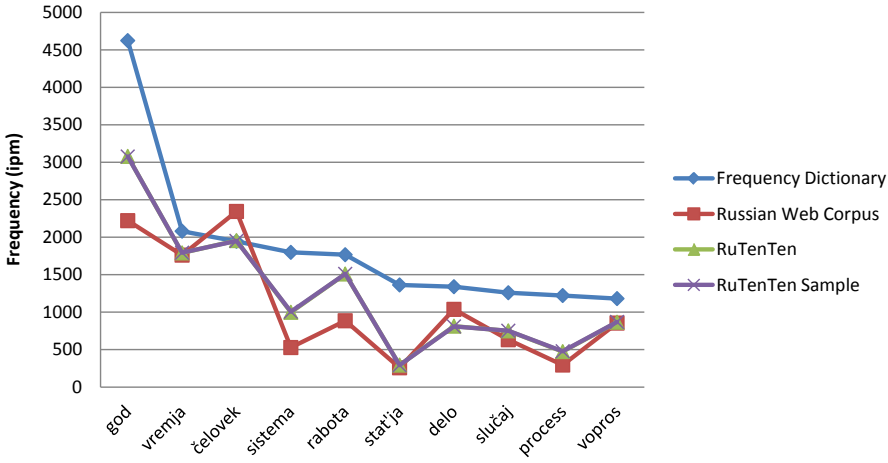


Fig. 1. Frequency distribution of nouns on the non-fiction word list (journalism excluded) as per three corpora (*y*-axis: frequency (ipm); charts: blue—Frequency Dictionary; red—Russian Web Corpus; green—ruTenTen; purple—ruTenTen Sample)

Table 6. Frequencies of nouns on the social and political journalism word list as per the three corpora

No.	Lemma	Translation	Frequency (ipm)			
			Social & political journalism word list in the Frequency Dictionary	Russian Web Corpus	ruTenTen	
					Corpus	Sample
1	god	year	5,589.50	2,220.74	3,078.97	3,076.99
2	čelovek	man	2,950.10	2,343.79	1,955.40	1,950.79
3	vremja	time	2,364.60	1,761.06	1,790.84	1,793.41
4	žizn'	life	1,548.40	1,054.70	864.40	862.25
5	delo	affair	1,482.00	1,037.09	813.12	809.29
6	den'	day	1,397.80	1,052.35	1,089.16	1,088.18
7	rabota	job	1,272.40	885.02	1,509.37	1,510.41
8	strana	country	1,203.90	576.81	662.05	664.03
9	vopros	question	992.00	853.94	866.03	869.27
10	slovo	word	989.70	807.83	633.83	631.81

On Fig. 2 we can see the data for the nouns in Table 2; like the results on Fig. 1 it shows that both the ruTenTen corpus and subset yield identical results. The word *rabota* (see Table 6) has higher frequency in the ruTenTen corpus, than in the Dictionary; for other nouns the Dictionary shows maximum frequency values. Spearman's rank correlation

coefficient between the ranked word lists in the Frequency Dictionary and in Russian Web Corpus is remarkably high standing at 0.94 which can indicate that Russian Web Corpus has more in common with newspaper articles. Only two nouns *vremja* and *den'* have identical rankings in Russian Web Corpus and the ruTenTen corpus.

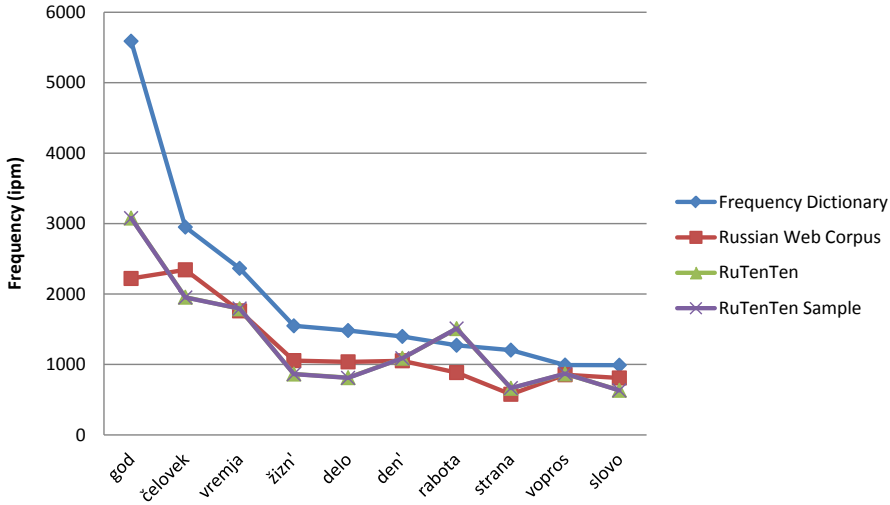


Fig. 2. Frequency Distribution of Nouns on the Social and Political Journalism Word List as per Three Corpora (*y*-axis: frequency (ipm); charts: blue—Frequency Dictionary; red—Russian Web Corpus; green—ruTenTen; purple—ruTenTen Sample)

At the next stage of our experiment we looked at the frequencies of nouns on the style-specific word lists (Tables 3 and 4). The results for the two style-specific groups of nouns are summarized in Table 7 and Fig. 3 and in Table 8 and Fig. 4 respectively.

It can be seen from Fig. 3 and 4 that the Frequency Dictionary subcorpus shows highest frequencies (in ipm), which is hardly surprising. Although style-specific words were selected not by absolute frequencies, but rather by the LL-score, this measure is still indicative of the number of units in the subcorpus, which in turn explains the fact why some lexemes with frequencies above corpus average values are marked as style-specific. The four nouns *sistema*, *process, sud, federacija*, *risunok* and *virus* have no discrepancy in ranking across the two corpora—Russian Web Corpus and ruTenTen. This is the largest number of words with identical ranking across the corpora. Spearman's rank correlation coefficient between the ranked word lists in the Frequency Dictionary and Russian Web Corpus is 0.92 and can indicate that the data on the non-fiction word list of style-specific vocabulary and the corpus data are largely identical (though to a lesser extent, the same is true for the ruTenTen corpus data, with the equally high Spearman's correlation coefficient between the Frequency Dictionary and the corpus standing at 0.73).

Table 7. Frequencies of style-specific nouns on non-fiction word list as per three corpora

No.	Lemma	Translation	Frequency (ipm)				
			Style-specific word list for non-fiction texts (journalism excluded) in the Frequency Dictionary		Russian Web Corpus	ruTenTen	
			Corpus	Subcorpus		Corpus	Sample
1	sistema	system	617.8	1,798.0	527.61	998.41	1,006.66
2	stat'ja	article	395.0	1,363.0	257.55	293.7	292.09
3	process	process	371.7	1,221.8	294.37	473.94	478.05
4	sud	court	371.1	1,153.9	197.00	302.98	301.73
5	federacija	federation	258.9	1,003.1	88.03	198.31	197.06
6	issledovanie	study	200.5	799.6	154.44	265.11	266.86
7	metod	method	197.0	772.3	153.51	263.74	265.91
8	ispol'zovanie	usage	190.3	757.9	146.83	346.74	350.04
9	risunok	picture	179.2	776.2	45.19	77.04	76.84
10	virus	virus	106.5	584.1	21.01	36.90	36.75

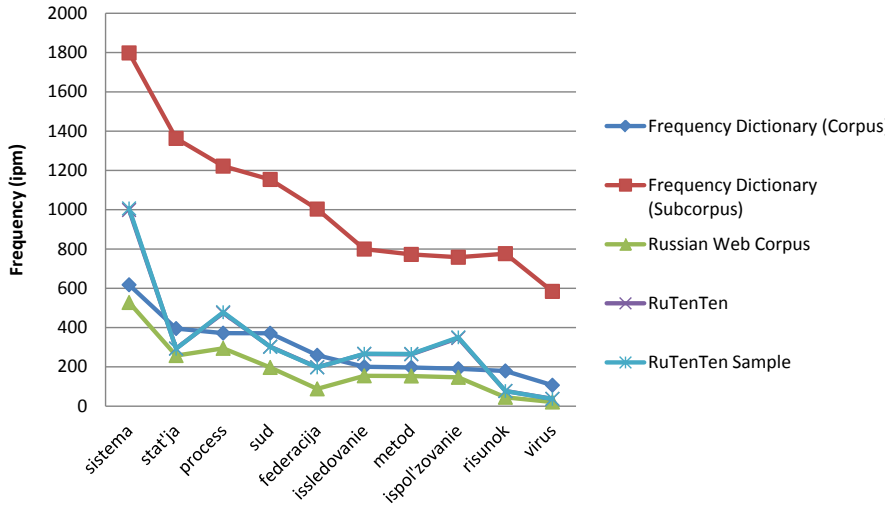


Fig. 3. Frequency distribution of style-specific nouns on the non-fiction word list as per three corpora (***y***-axis: frequency (ipm); charts: blue—Frequency Dictionary (corpus); red—Frequency Dictionary (subcorpus); green—Russian Web Corpus; purple—ruTenTen; light blue—ruTenTen Sample)

Table 8. Frequencies of nouns on the style-specific word list for news and newspaper texts as per three corpora

No.	Lemma	Translation	Frequency (ipm)				
			Style-specific word list for news and newspaper texts from the Frequency Dictionary		Russian Web Corpus	ruTenTen	
			Corpus	Subcorpus		Corpus	Sample
1	god	year	3,727.5	5,589.5	2,220.74	3,078.97	3,076.99
2	strana	country	725.7	1,203.9	576.81	662.05	664.03
3	kompanija	company	392.7	699.0	390.72	970.15	979.11
4	prezident	president	311.0	634.6	185.6	215.07	213.81
5	teatr	theatre	305.3	611.9	91.08	102.09	99.28
6	pravi-tel'stvo	government	277.7	531.2	183.25	225.28	224.83
7	fil'm	film	196.8	380.1	172.15	214.16	213.10
8	spektakl'	play	164.7	350.0	37.09	44.42	42.78
9	reforma	reform	133.1	273.0	58.48	47.16	47.74
10	vybory	elections	117.7	243.4	—	62.34	63.20

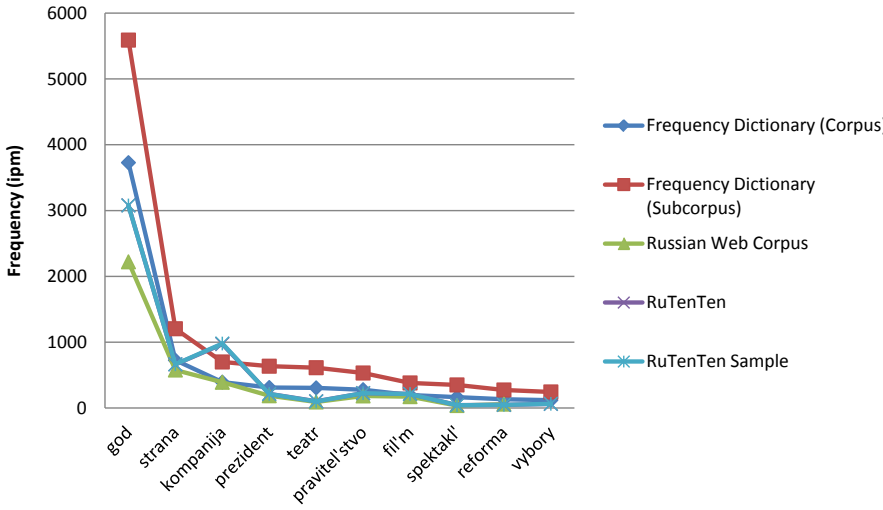


Fig. 4. Frequency distribution for nouns on the style-specific word list for news and newspaper texts as per three corpora (*y*-axis: frequency (ipm); charts: blue—Frequency Dictionary (corpus); red—Frequency Dictionary (subcorpus); green—Russian Web Corpus; purple—ruTenTen; light blue—ruTenTen Sample)

It is particularly remarkable, that the maximum initial value on the chart corresponds to the frequency of *god* and is present on each graph. The explanation is that this noun is top three by frequency in all the three corpora. The Russian Web Corpus failed to produce any results for the lexeme *vybory*, because according to its morphological token the usage of this lemma merges with the lemma *vybor*. The first four nouns *god, strana, kompanija* and *prezident* have identical ranking in the Russian Web Corpus and the Frequency Dictionary (both in the main corpus and the subcorpus). Spearman's rank correlation coefficient is equally top high for the Frequency Dictionary and Russian Web Corpus standing at 0.95.

2.3. Low Frequency Nouns: Selection

The selection of low-frequency nouns is quite tricky. For this aim we analyzed the Frequency Dictionary paying attention to the tail of the frequency list and selected 12 low-frequency nouns ranked between the following two ranges: 1) 18,940–18,965; 2) 19,955–20,000 in the Frequency Dictionary [Lyashevskaya. Sharoff 2009].

Table 9. Low-frequency nouns in the Frequency Dictionary[2]

No.	Lemma	Translation	Frequency indicated in the Dictionary (ipm)[2]	Frequency (absolute)			Frequency (ipm)		
				Russian Web Corpus	ruTenTen	ruTenTen Sample	Russian Web Corpus	ruTenTen	ruTenTen Sample
1	opala	disgrace	2.6 (4.3)	200	15,235	1,067	1.06	0.83	0.85
2	zador	fervour	2.8 (3.7)	322	26,173	1,826	1.71	1.43	1.46
3	svastika	swastika	2.6 (3.6)	496	29,270	1,902	2.64	1.60	1.52
4	šljuz	sluice	2.6 (2.9)	851	87,209	5,867	4.53	4.77	4.68
5	sčastlivec	lucky man	2.6 (2.6)	262	12,256	856	1.39	0.67	0.68
6	zlodejstvo	outrage	2.8 (2.5)	332	13,768	958	1.77	0.75	0.76
7	inkvizicija	inquisition	2.6 (2.5)	552	48,051	3,168	2.94	2.63	2.53
8	zaplata	patch	2.6 (1.9)	325	17,078	1,182	1.73	0.93	0.94
9	xoluj	groveller	2.6 (1.7)	195	8,453	540	1.04	0.46	0.43
10	tjulen'	seal	2.6 (1.3)	375	29,725	1,776	2.00	1.63	1.42
11	zagrivok	nape	2.8 (1.0)	225	9,695	655	1.20	0.53	0.52
12	sedmica	week	2.6 (0.8)	202	20,628	1,722	1.07	1.13	1.37

[2] Frequency indicated in social and political journalism list (1990–2000s) is written in parentheses.

2.4. Low Frequency Nouns: Results

The data shown in Table 9 suggests that the differences between the Frequency Dictionary and corpora are not apparently reconciled in case of low frequency words. Spearman's coefficient has the following values (as compared to the results with the high-frequency nouns): 0.33 for the Frequency Dictionary and ruTenTen sample corpus, 0.22 for the Frequency Dictionary and ruTenTen corpus and 0.21 for the Frequency Dictionary and the Russian Web corpus. However, the Russian Web Corpus shows similar high coherence in data both with the ruTenTen and ruTenTen sample corpora (Spearman's correlation coefficient is 0.80 and 0.79 respectively). Hence for the given low-frequency words the difference between the corpora and the Frequency Dictionary is more obvious than between the corpora in question. This fact can imply that corpora automatically crawled from the web share more in common with each other that with the National Russian Corpus (that served as the source for the Frequency Dictionary). Low values of the Spearman's coefficient imply that nouns differ a lot in their ranks between the latter corpus and other corpora and it is crucial in case of low-frequency words.

3. Conclusion and Future Work

The general conclusion from the obtained data suggests that texts selected for large corpora reflect the language of the web. The results published in the Frequency Dictionary are based on the Russian National Corpus, which makes them so coherent. High-frequency nouns indicated in non-fiction texts tend to be more similar to the ruTenTen corpus, whereas words fixed in the social and political journalisms subcorpus share a lot with data in the Russian Web Corpus. Thus, the Russian Web Corpus has more in common with newspaper articles. The analysis of high-frequency nouns and their ranking positions in both 1 bln subset and 14 bln corpus shows that they have produced the same results, but this is not true for the low frequency nouns. In case of low frequency data three corpora do not show much coincidence with the Frequency Dictionary lists. However we can say that in general the sample shares similar features with the total set (ruTenTen) and hence in this sense small corpora can be used for evaluating word frequencies.

The nouns (*sajt*, *sistema, kompanija,* and *Rossija*) that are ranked top by frequency in the ruTenTen corpus and its billion-size subset, but are missing among the results in the Frequency Dictionary, reveal the specific properties of web-based texts—firstly, their abundance and secondly, the focus on describing the web-page content. If we use more high-frequency nouns the Spearman's coefficient will be lower because of the diversity in ranks of the words. But the value of the coefficient will be constantly higher (than if we increase the number low-frequency words as it will be even negative).

The Russian Web Corpus appears to be more consistent with the Frequency Dictionary than the ruTenTen corpus in describing high-frequency nouns. The differences between the corpora are apparently reconciled in case of high-frequency words, but the opposite doesn't hold true for the low-frequency words.

We believe that these experiments have a future. It is crucial to study the results for low frequency words, because this group of words is the one that may produce entirely different numeric values for large corpora. To be more specific, preliminary results of our collocations study have shown that higher absolute frequency of a particular lexical item is not always conducive to a larger number of relations for the said item (despite greater number of syntagmatic partners, typical for each relation).

Acknowledgements

This work was supported by the grant of the President of Russian Federation for state support of scholarly research by young scholars (Project No. MK-5274.2016.6).

References

1. *Baroni M., Bernardini S.* (2004), BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC 2004, Lisbon: ELDA, pp. 1313–1316
2. *Belikov V. I., Selegey V. P., Sharoff S. A.* (2012) Preliminary considerations towards developing the General Internet Corpus of Russian [Prolegomeny k proyektu General'nogo internet-korpusa russkogo yazyka (GIKRYa)]. In Computational linguistics and intellectual technologies. Vol. 11 (18). Moscow: Izd-vo RGGU, pp. 37–49.
3. *Benko V.* (2014), Aranea: Yet Another Family of (Comparable) Web Corpora. In: P. Sojka, A. Horák, I. Kopeček and K. Pala (Eds.): Text. Speech and Dialogue. 17th International Conference. TSD 2014. Brno. Czech Republic. September 8–12 2014. Springer International Publishing, pp. 257–264.
4. *Shu-Kai Hsieh* (2014), Why Chinese Web-as-Corpus is Wacky? Or: How Big Data is Killing Chinese Corpus Linguistics? In Proceedings of the 9th Edition of the Language Resources and Evaluation. Reykjavik, Iceland, pp. 2386–2389.
5. *Jakubíček M., Kilgarriff A., Kovář V., Rychlý P., Suchomel V.* (2013), The TenTen Corpus Family. In Proceedings of the International Conference on Corpus Linguistics, pp. 125–127.
6. *Kehoe A., Renouf A.* (2002), WebCorp: Applying the Web to Linguistics and Linguistics to the Web. In Proceedings WWW2002 Conference. Honolulu, Hawaii, available at: http://www2002.org/CDROM/poster/67.
7. *Kilgarriff, A.* (2001), Comparing corpora. In International journal of corpus linguistics. 6(1), pp. 97–133.
8. *Kilgarriff A., Grefenstette G.* (2003), Introduction to the Special Issue on Web as Corpus. In Computational Linguistics, 29 (3), pp. 333–347.
9. *Lyashevskaya O., Sharoff S.* (2009), Frequency Dictionary of Contemporary Russian based on the Russian National Corpus data [Chastotnyj slovar' sovremennogo russkogo jazyka (na materialakh Natsional'nogo Korpusa Russkogo Jazyka)]. Moscow: Azbukovnik.
10. *Piperski A., Belikov V., Kopylov N., Morozov Eu., Selegey V., Sharoff S.* (2013), Big and diverse is beautiful: A large corpus of Russian to study linguistic variation.

In Proceedings of the 8th Web as Corpus Workshop (WAC-8) Corpus Linguistics Conference 2013, available at: https://sigwac.org.uk/raw-attachment/wiki/WAC8/wac8-proceedings.pdf

11. *Shaikevich A. Ya.* (2015), Measures of Lexical Similarity between Frequency Dictionaries [Mery leksičeskogo sxodstva častotnyx slovarej], Proc. International Conference "Corpus linguistics-2015" [Trudy Mezhdunarodnoy konferencii "Korpusnaya lingvistika–2015"], St. Petersburg: St. Petersburg State University, pp. 434–442.
12. *Sharoff S.* (2006), Creating general-purpose corpora using automated search engine queries. In Marco Baroni and Silvia Bernardini. (eds). WaCky! Working papers on the Web as Corpus. Gedit, Bologna, available at: http://wackybook.sslmit.unibo.it/
13. *Sinclair J.* (2004), Trust the text: Language, corpus and discourse. London/New York: Routledge.
14. *Zakharov V.* (2015), Evaluation of Internet corpora of Russian [Ocenka kačestva Internet-korpusov russkogo jazyka], Proc. International Conference "Corpus linguistics-2015" [Trudy Mezhdunarodnoy konferencii "Korpusnaya lingvistika–2015"], St. Petersburg: St. Petersburg State University, pp. 219–229.

КОАРТИКУЛЯЦИЯ НА СТЫКАХ СЛОВ КАК ПОКАЗАТЕЛЬ НАЛИЧИЯ ПРОСОДИЧЕСКОГО ШВА В РУССКОМ ЯЗЫКЕ[1]

Князев С. В. (svknia@gmail.com)

МГУ им. М. В. Ломоносова; НИУ ВШЭ, Москва, Россия

Ключевые слова: фонетика, просодическое членение, просодический шов, коартикуляция, паузальный маркер, инструментальный анализ

VOICE COARTICULATION ACROSS WORD BOUNDARIES AS A CUE FOR DETECTING PROSODIC BREAKS IN STANDARD RUSSIAN

Knyazev S. V. (svknia@gmail.com)

Moscow State University; Higher School of Economics, Moscow, Russia

The paper reports some results of the research, aimed at finding out whether regressive and / or progressive voice coarticulation available in clusters of homorganic labiodental consonants /v/ + /v/ in an external sandhi position in Modern Standard Russian may serve as a cue for detecting the location and depth of prosodic breaks.

Combinations of labiodental fricatives /v/ + /v/ at the word junctures result in [ff], [vv] or [fv] pronunciation (with the decreasing abundance) in Modern Standard Russian. The percentage ratio of the above mentioned pronunciation types depends on the strength of the prosodic break between two words:

- in the position within an intonation group (no prosodic break) [ff] pronunciation appears fairly stable and makes about 70% of the total case number, while the percentage of [fv] pronunciation (corresponding to the absence of the coarticulation) varies in the range of 1–11%.
- in the position around prosodic break between two words group [fv] pronunciation detected in more then 80% out of the total case number studied.

Key words: phonetics, prosodic phrasing, prosodic breaks, word boundary strength, pause marker, coarticulation, instrumental analysis

[1] Исследование проведено при поддержке гранта РФФИ 15-06-06103.

0. В настоящем докладе представлены результаты экспериментально-фонетического исследования уподобления согласных по голосу на стыках фонетических слов (на материале губно-зубных спирантов), На основании полученных данных обсуждается вопрос о том, может ли наличие и/или характер этого уподобления служить показателем наличия или отсутствия просодической границы в звучащем тексте.

1. Просодическое членение звучащего текста — это его разбиение на супрасегментные единицы при помощи звуковых средств (на фразовом уровне — на синтагмы. Границы между фразовыми просодическими составляющими образуют просодические швы («prosodic breaks»), абстрактные показатели членения между просодическими составляющими, имеющие то или иное фонетическое воплощение (часто вариативное), а также определенные перцептивные корреляты. Таким образом, просодический шов — это абстрактный показатель сегментирующего потенциала словораздела, который реализуется в тексте с разной вероятностью и с разной сегментирующей силой [Кривнова 2015].

В отечественной лингвистике впервые обратил внимание на просодическое членение и его иерархическую природу академик Л. В. Щерба [Щерба 1915]; он отмечал также, что просодическое членение зависит от стиля и жанра текста, ср. [Светозарова 2000]. В литературе описывается и корреляция между типом устной речи (чтение или спонтанная речь) и членением предложений на синтагмы: в спонтанной речи просодические синтагмы короче, чем при чтении (и составляют в среднем 2,7 слова против 4,2 слова) [Иомдин, Лобанов 2009]. Очевидно, что характер просодического членения зависит от идиолекта [Светозарова 1982]. Существует также связь между просодическим членением и эмоциональным состоянием говорящего, а также индивидуальным темпом речи, установкой на степень выразительности речи или чтения. Как показали результаты проведенных экспериментов, проблема паузирования тесно связана с более общей проблемой вариативности манер чтения[2] [Брызгунова 1980: 98–99].

О. Ф. Кривнова описывает 4-х балльную шкалу глубины просодического шва, которые соответствуют просодическим составляющим, большим, чем фонетическое слово, следующим образом: просодическая/ фонетическая синтагма; интонационная фраза; интонационно-смысловой комплекс; высказывание [Кривнова 1995].

Несмотря на интересное и продуктивное обсуждение иерархической природы просодического членения и контролирующих его факторов в литературе XX в., до 80–90-х годов конкретных исследований просодического членения в речи было очень мало как на материале русского, так и на материале других языков. В 80–90-е гг. в фонетике и в лингвистике в целом произошел «просодический бум», тесно связанный с переходом от преимущественно структурного подхода к функциональной и когнитивной научной парадигме, изучению

---

[2] Подробно проблема произносительной нормы в области интонации в связи с просодическим членением рассматривается в [Кривнова, Чардин 1999].

устного дискурса, «языка в действии», интересом к компьютерным моделям языка и устной речи, к разработкам по автоматическому синтезу речи, невозможному без понимания функций и природы просодического членения. Основные направления его исследования в современной лингвистике сформулированы О. Ф. Кривновой следующим образом: «главные направления исследований ПЧ[3] (теоретических, экспериментально-инструментальных, прикладных) группируются вокруг следующих проблем:

- Локальные маркеры (границы, просодические швы) ПЧ — текстовая локализация, глубина создаваемого членения (сегментирующая сила границ, их иерархия), средства фонетической реализации.
- Квантованная/блочная природа просодических составляющих, их иерархический статус, интегрирующие просодические схемы разного уровня, их фонетическая реализация.
- Функциональный аспект ПЧ, контролирующие факторы: коммуникативные, семантико-синтаксические, психофизиологические (когнитивные, речепроизводящие)» [Кривнова 2015].

В конкретных исследованиях просодические швы могут анализироваться при интерпретации уже озвученного текста, то есть, с позиции слушающего, который воспринимает текст, или с позиции говорящего, который порождает или читает письменный текст. Центральной же остается задача выявления факторов, контролирующих просодическое членение, и описание их просодической реализации, что актуально как для фонетической науки, так и для разработки прикладных речевых систем.

В существующей русскоязычной литературе анализируются различные фонетические и нефонетические корреляты просодических швов, в основном на материале работ по автоматическому синтезу речи, для которого необходимы эксплицитные алгоритмы задания и акустической реализации просодического членения.

1.1. Что касается **нефонетических коррелятов просодических швов** в тексте, то, согласно имеющимся экспериментальным данным, в письменном тексте границы самых крупных просодических составляющих однозначно коррелируют с границами абзацев и при этом всегда оформляются при помощи физических пауз и других просодических средств [Светозарова 2000, 120]; [Чардин 1999, 66]. Первичное грубое членение текста при синтезе речи обычно опирается на знаки препинания. Существенную роль для предсказания наличия или отсутствия просодических швов может играть расстояние между просодическими границами в графических и фонетических словах, и длительность предыдущего и последующего отрезков относительно данной точки [Светозарова 2000], [Цирульник, Лобанов, Сизонов 2008], [Хомицевич, Соломенник 2010].

Традиционно считается, что существенным фактором, влияющим на ПЧ, является дыхательный ритм: средняя частота дыхательных пауз в речи составляет

[3] Просодического членения

16–20 в минуту [Потапова, Блохина 1986], следовательно в среднем после отрезка речи длиной 3–3,75 секунд ожидается просодический шов.

При синтезе речи алгоритмы просодического членения при отсутствии знаков препинания в длинном текстовом фрагменте определяют количество фонетических слов в нем. Длина последовательности знаменательных / фонетических слов, внутри которой постулируется необходимость просодической границы, определяется разными исследователями по-разному [Светозарова 2000 с. 125]; [Цирульник, Лобанов, Сизонов 2008]; [Хомицевич, Соломенник 2010].

Аналогично знакам препинания некоторые лексемы также задают границы просодических синтагм и фраз или, наоборот, их отсутствие. Последнее характерно для служебных / функциональных слов, которые образуют закрытый список и к тому же очень частотны [Светозарова 1982: 125]. Однако, в целом, принадлежность слова к определенной части речи влияет на вероятность появления после него просодического шва в значительно меньшей степени, чем наличие знака препинания на словоразделе [Лобанов, Гецевич 2011].

1.2. В литературе выделяются следующие **фонетические параметры**, которые значимы для слушающего и могут быть использованы при детектировании и идентификации просодических швов в звучащем тексте и их акустической реализации в случае автоматического синтеза речи:

- наличие физических пауз, «в том числе абсолютных и заполненных», а также их длительность[4];
- синтаксические акценты, прежде всего тональные (между двумя такими акцентами обычно наличие просодического шва);
- некоторые значимые движения тона за пределами акцентов, в частности, пограничные тоны;
- квазисегментные явления (ларингализация, придыхание) и элементы речевого дыхания);
- изменение тонального регистра (возвращение на базовый уровень);
- замедление темпа произнесения перед просодическим швом (финальное продление), ускорение после него;
- уменьшение интенсивности перед просодическим швом, увеличение после;
- эмфатическая просодия;

---

[4] Темпоральные (физические) паузы выделяются в качестве наиболее очевидного пограничного сигнала (подробный обзор вопросов, связанных с паузированием при автоматическом синтезе речи, см. [Чардин 1999]). Паузы подразделяются на грамматические (имеющие отношение к смыслу и синтаксису предложения) и неграмматические (хезитационные и выделительные) [Светозарова 1982]. Собственно паузы (темпоральные или чем-то заполненные) не являются обязательным коррелятом границы просодических синтагм, но очень вероятны на границах предложений [Коротаев 2009]. Вероятность наличия паузы в качестве средства реализации просодического шва повышается при произнесении предшествующих просодических синтагм без пауз между ними [Коротаев 2009]. Особую группу физических пауз представляют собой **дыхательные паузы (ДП)** — это интонационно-смысловые паузы с включенным в них вдохом (при этом физиологически обусловлена необходимость включения вдоха в некоторые темпоральные интонационные паузы, но не сами по себе дыхательные паузы).

- **особенности фонетической реализации фонем** (например, отсутствие редукции в конечных открытых слогах) [Кибрик, Кодзасов, Худякова 2009].

2. Данная работа посвящена исследованию одного частного случая фонетической реализации фонем: в ней анализируется явление коартикуляции на стыках фонетических слов как один из способов маркировки / детектирования просодических границ в связном тексте при отсутствии физической паузы.

2.1. Одним из очевидных случаев подобной маркировки является характер реализации глухих/звонких согласных на стыках слов — так, отсутствие **ассимиляции** по глухости/звонкости, то есть, наличие глухого шумного согласного в положении перед звонким шумным (*Иваново[ф] давно уехал*) явным образом маркирует наличие просодической границы. Гораздо меньше известно о том, как устроена в подобных случаях **коартикуляция** согласных по тому же признаку.

В современном русском литературном языке согласный <в> (и <в’>), «в зависимости от характера последующего звука функционирует либо как сонант (<в> + гласный, <в> + сонорный), либо как звонкий шумный (<в>+звонкий шумный)» [Пауфошима 1969: 150]; иначе говоря, для правил, регулирующих реализацию глухих/звонких согласных, он выступает как «пустое место», «прозрачный», нулевой сегмент, поскольку произношение глухого или звонкого согласного перед ним зависит от того, какой сегмент следует за ним [Jacobson 1956: 98].

В соответствии с нормами СРЛЯ, можно предполагать, что звонкие губные согласные на конце слова перед <в> (и <в’>) следующего слова должны быть реализованы своими глухими аллофонами, как и все другие звонкие шумные: в литературе постулируется произношение типа *пло*[ф#в]*ышел* в соответствии с несомненным *пру*[т#в]*ысох* [Пауфошима 1969: 150]. Однако как слуховой, так и инструментальный анализ показывают, что на стыках слов сочетания из гоморганных согласных, вторым из которых является <в>/<в’>, часто произносятся не так, как сочетания согласных разного места образования (особенно в тех случаях, когда эти согласные совпадают и по способу артикуляции). Так, если в случаях типа *друг Васи*, *груз выслан* и т. п. в литературном русском языке на месте <г> и <з> перед <в>/<в’> обязательно произношение глухого согласного [к], [с] и т. п. (т. е. [дру́к ва́с’и], [гру́с вы́слън]), то на месте <ф#в> и <в#в> (<ф#в’> и <в#в’>) при отсутствии просодического шва фиксируется произношение звонкого долгого [в] или сочетания, близкого к [фф][5], например, *в семь часов вечера* [ч’исо́в в’е́ч’ьра] / [ч’исо́ф в̯’е́ч’ьра], *часов восемь утра* [ч’исо́в во́с’ьм’] / [ч’исо́ф в̯о́с’ьм’], *напротив ворот* [напро́т’ьв варо́т] / [напро́т’ьф в̯аро́т], *Петров-Водкин* [п’итро́в во́тк’ьн] / [п’итро́ф в̯ о́тк’ьн], что может быть объяснено наличием кортикуляции по голосу [Князев 1999], [Knyazev, Vorontsova, Petrova 2007].

---

[5] Следует, впрочем, отметить, что второй согласный в этом случае отличается от «настоящего» [ф] меньшей интенсивностью шума, то есть, представляет собой «глухой [в]» (который было бы точнее обозначать в транскрипции как [в̯]).

На то, что данное явление следует характеризовать как коартикуляцию, а не как ассимиляцию, указывает именно характер прогрессивного оглушения — это уподобление не полное (не по всем параметрам дифференциального признака глухость / звонкость), а лишь по наличию / отсутствию фонации (характер же шума остается у оглушенного звонкого тем же): см. рис. 1, на котором приведены осциллограммы и динамические спектрограммы сочетаний *Петров-Водкин* (с произношением [вв]), *Петров-Водкин* ([фв̯]) и *Петров-Фоткин* ([фф]).

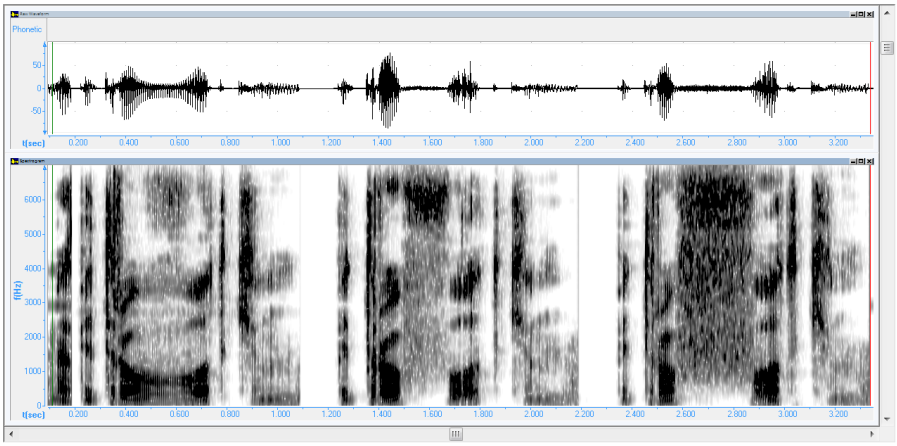


Рисунок 1. Слева направо — осциллограммы и динамические спектрограммы сочетаний *Петров-Водкин* ([вв]), *Петров-Водкин* ([фв̯]) и *Петров-Фоткин* ([фф])

2.2. Предыдущее исследование [Knyazev, Vorontsova, Petrova 2007] было посвящено анализу реализации указанных консонантных сочетаний **внутри синтагмы** (при отустствии просодических швов между тестовыми словами); его результаты свидетельствуют о том, что реализация эта, тем не менее, зависит от характера просодического оформления высказывания: при наличии тонального акцента на одном из слов, особенно на первом, увеличивается количество реализаций [ф#в] за счет других возможных вариантов, в первую очередь, за счет [в#в] (с 2% до почти 20% от общего числа[6] произнесений).

2.3. Целью настоящего исследования была проверка гипотезы о том, что характер реализации согласных на месте глубинного сочетания <в#в> как [вв], [фф] или [фв] зависит от наличия/отсутствия просодической границы между словами, в которых находятся данные фонемы.

---

[6] Подробно проблема произносительной нормы в области интонации в связи с просодическим членением рассматривается в [Кривнова, Чардин 1999].

В качестве дикторов в эксперименте приняли участие 78 **информантов** (преимущественно женщины в возрасте 17–20 лет), носители современного литературного произношения.

Материалом исследования служили 4 тестовых предложения с сочетанием слов *плов варится* без физической паузы между ними; варьируемым параметром в которых была просодическая организация высказывания. Ниже под номером 1 приведены примеры, повторяющие материал предыдущего эксперимента, под номером 2 — новые примеры, являвшиеся основным материалом данного эксперимента:

1.1 наличие тонального акцента на слове *плов*, отсутствие его на слове *варится* (*Но учтите, это именно пло́в варится, а овощи лучше обжарить*);
1.2 отсутствие тонального акцента на слове *плов*, наличие его на слове *варится* (*Когда плов ва́рится, крышку лучше не поднимать*);
1.3 отсутствие тонального акцента на словах *плов* и *варится* (*Если плов варится дольше, получается просто каша*);
1. наличие тонального акцента на словах *плов* и *варится* (*Мясо сначала тушится, а пло́в ва́рится*)
при отсутствии знаков препинания между ними.

Предложения были зачитаны информантами вслух в составе связного текста, представлявшего собой сконструированный рецепт приготовления плова, и записаны непосредственно в память компьютера.

Анализ производился при помощи программ *PRAAT* и *Speech Analyzer*, в его ходе была измерена длительность глухих и звонких участков на отрезке реализации губно-зубных спирантов в сочетании *плов варится*.

Алгоритм принятия решения о характере реализации консонантного кластера был следующим:

- полное отсутствие глухого участка свидельствовало о реализации сочетания как [вв];
- полное отсутствие звонкого участка или наличие звонкого участка длительностью не более 25 мс в начале или конце консонантного отрезка считалось реализацией [фф] (точнее, [фв̥]);
- наличие звонкого участка длительностью более 25 мс в конце консонантного отрезка свидельствовало о реализации [фв];

см. ниже на рис. 2–4 примеры произнесения сочетания *плов варится* во фразах

1.1 *Но учтите, это именно плов варится, а овощи лучше обжарить* ([фв̥]);
1.3 *Если плов варится дольше, получается просто каша* ([вв]);
2 *Мясо сначала тушится, а плов варится* ([фв]).

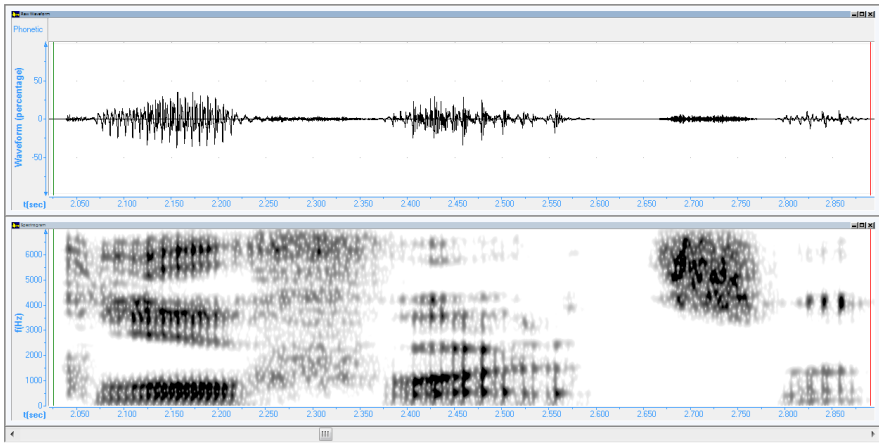


Рисунок 2. Осциллограмма и динамическая спектрограмма сочетания *плов варится* из фразы 1.1 *Но учтите, это именно плов варится, а овощи лучше обжарить* (произношение [фв̯])

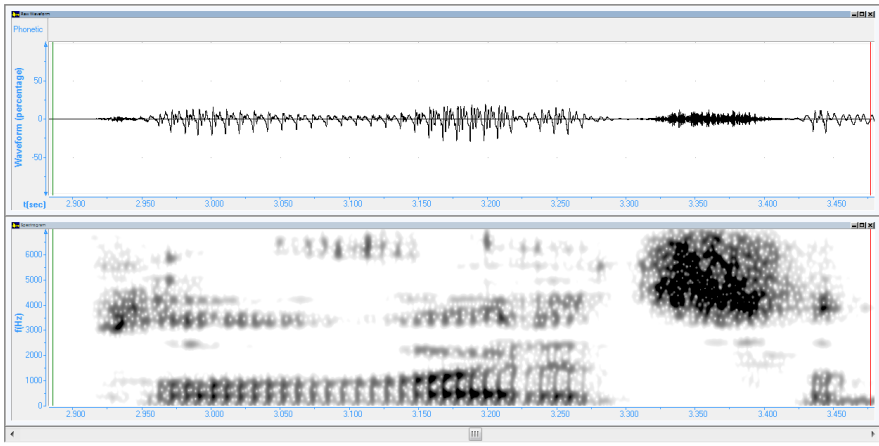


Рисунок 3. Осциллограмма и динамическая спектрограмма сочетания *плов варится* из фразы 1.3 *Если плов варится дольше, получается просто каша* (произношение [вв])

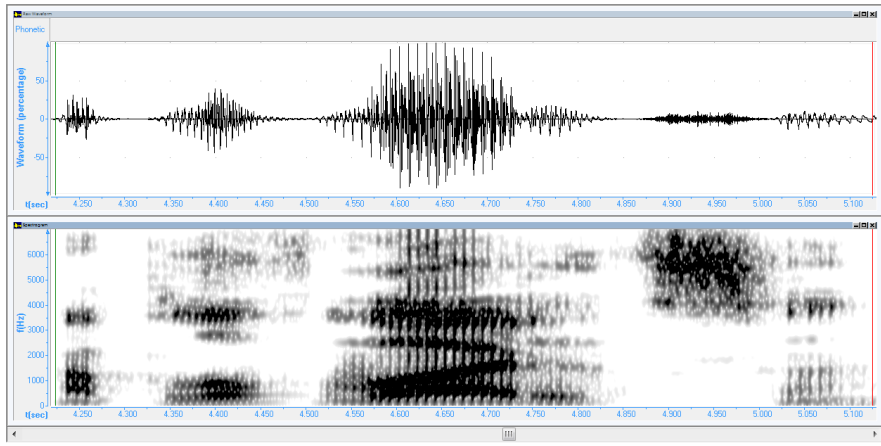


Рисунок 4. Осциллограмма и динамическая спектрограмма сочетания *а плов варится* из фразы 2 *Мясо сначала тушится, а плов варится* (произношение [фв])

В ходе анализа результатов из рассмотрения были исключены те прочтения, в которых фиксировались темпоральные (физические) паузы. Таких случаев оказалось 11 из 322; бóльшая часть из них — в предложении № 2 (6 из 78), остальные 5 в других произнесениях тех же дикторов из числа шести, реализовавших паузу в предложении № 2. Все другие данные этих дикторов были также исключены из рассмотрения, таким образом, окончательное число испытуемых, от которых был получен материал, проанализированный в ходе настоящего исследования, составило 72, общее число обработанных произнесений — 288.

Результаты измерения длительности глухого // звонкого участка аллофонов фонем <в#в> при отсутствии физической паузы между ними для всех дикторов во всех фразах обобщены в Таблице 1. В таблице сначала приводится количество реализаций, затем в скобках — процент от их общего числа.

Таблица 1. Количество произнесений [фв] [вв] и [фв̯] на месте сочетания <в#в> во фразах

1.1 (*Но учтите, это именно плов варится, а овощи лучше обжарить*);
1.2 (*Когда плов варится, крышку лучше не поднимать*);
1.3 (*Если плов варится дольше, получается просто каша*);
2 (*Мясо сначала тушится, а плов варится*)

при отсутствии физической паузы между ними

кол-во / фраза	[фв]	(%)	[вв]	(%)	[фв̯] = [фф]	(%)	всего	
1.1	8	(11%)	8	(11%)	56	(78%)	72	(100%)
1.2	3	(4%)	27	(37%)	42	(59%)	72	(100%)
1.3	1	(1%)	37	(53%)	34	(46%)	72	(100%)
2	58	(81%)	3	(4%)	11	(15%)	72	(100%)

Приведенные в таблице данные[7] свидетельствуют о том, что произношение [фв] (без коартикуляции) при наличии просодического шва наблюдается в 81 % всех проанализированных реплик (58 из 72), а при его отсутствии — лишь в 5 % (12 из 216). В последнем случае наличие акцента, особенно на первом слове, несколько повышает верятность отсутствия коартикуляции.

Таким образом, на основании полученных результатов можно сформулировать следующий **вывод**:

- отсутствие коартикуляции согласных на месте сочетания <в#в> на стыках фонетических слов может служить достаточно надежным показателем наличия просодической границы между этими словами, а наличие взаимодействия согласных по голосу в этом положении — показателем отсутствия просодического шва.

Литература

1. *Аванесов Р. И.* Русское литературное произношение. 6-е изд. М., 1984.
2. *Брызгунова Е. А.* Интонация // Русская грамматика. Т. 1. М., 1980.
3. *Иомдин Л. Л., Лобанов Б. М.* Синтаксические корреляты просодически маркированных элементов предложения и их роль в задачах синтеза речи по тексту // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». М., 2009.
4. *Кибрик А. А., Кодзасов С. В., Худякова М. В.* Просодическая транскрипция: уровни детализации // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». М., 2009.
5. *Князев С. В.* О мягкости необычайной // Фонетика и грамматика: настоящее, прошедшее, будущее: к 50-летию научной деятельности С. К. Пожарицкой. — Т. 13 из Вопросы русского языкознания. — Изд-во МГУ Москва, 2010. — С. 61–70.
6. *Князев С. В.* О прогрессивной ассимиляции в современном русском языке // Вестник МГУ. Сер. 9. Филология. 1999, N 4.
7. *Коротаев Н. А.* Отсутствие пауз на границах элементарных дискурсивных единиц: опыт корпусного исследования // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». М., 2009.
8. *Кривнова О. Ф.* Глубина просодических швов в звучащем тексте (экспериментальные данные) // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Москва, 27–30 мая 2015 г.). Вып. 14 (21). — М.: Изд-во РГГУ, 2015.
9. *Кривнова О. Ф.* Перцептивная и смысловая значимость просодических швов в связном тексте // Проблемы фонетики II. С. 228–238. М., 1995.

---

[7] Некоторое их отличие от данных, полученных в [Knyazev, Vorontsova, Petrova 2007] (в сторону общего увеличения количества примеров с наличием коартикуляции) может быть объяснено возрастом информантов — меньшим в настоящем исследовании.

10. *Кривнова О. Ф., Чардин И. С.* Паузирование при автоматическом синтезе речи // Теория и практика речевых исследований (АРСО-99). М., 1999.
11. *Лобанов Б. М., Гецевич Ю. С.* Статистические характеристики синтагматического членения предложений в приложении к синтезу выразительной речи по тексту // Труды Международной конференции «Компьютерная лингвистика и интеллектуальные технологии» (Диалог'2011), Бекасово 25–29 мая 2011, Вып. 10 (17). — М.: РГГУ, 2011.
12. *Панов М. В.* Современный русский язык: Фонетика. М., 1979.
13. *Пауфошима Р. Ф.* Некоторые вопросы, связанные с категорией глухости/звонкости согласных в говорах русского языка // Экспериментально-фонетическое изучение русских говоров. М., 1969.
14. *Потапова Р. К., Блохина Л. П.* Средства фонетического членения речевого потока в немецком и русском языках. — М., 1986.
15. *Светозарова Н. Д.* Роль фразовой интонации в речевой деятельности и возможности ее моделирования // Фонология речевой деятельности. СПб.: Изд. С.-Петербургского ун-та, 2000.
16. *Светозарова Н. Д.* Интонационная система русского языка. — Л., 1982.
17. *Хомицевич О. Г., Соломенник М. В.* Автоматическая расстановка пауз в системе синтеза русской речи по тексту // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». М., 2010.
18. *Цирульник Л. И., Лобанов Б. М., Сизонов О. Г.* Алгоритм интонационной разметки повествовательных предложений для синтеза речи по тексту // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». М., 2008.
19. *Чардин И. С.* Проблема паузирования при автоматическом синтезе речи. Дипломная работа. М., 1999.
20. *Щерба Л. В.* Восточнолужицкое наречие. Пгр., 1915.
21. *Jacobson, Roman.* Die Verteilung der stimmhaften und stimmlosen Geräuschlaute in Russischen // Festschrift für Max Vasmer. Berlin. 1956.
22. *Knyazev S. V., I. I. Vorontsova & I. V. Petrova.* Voice Coarticulation Across Word Boundaries in /v/+/v/ Sequences in Standard Russian // AFCP Workshop «Coarticulation : cues, direction, and representation». Montpellier, France. December 7, 2007 (Workshop de l'association Francophone de la Communication Parlée La co-articulation: Indices, Direction et Représentation. Organisé conjointement avec les laboratories/ DIPRALANG (EA 739) et PRAXILING (UMR CNRS 5267).

References

1. *Avanesov R. I.* (1984) Russian Literary Pronunciation [Russkoe literaturnoe proiznoshenie], Education, M.
2. *Bryzgunova E. A.* (1980), Intonation [Intonaciya] // Russian grammar [Russkaya grammatika], vol. I.

3. *Chardin I. S.* (1999), The problem of pauses' placement for text-to-speech synthesis [Problema pauzirovaniya pri avtomaticheskom sinteze rechi]. Diploma paper. MGU, philological faculty.
4. *Iomdin L. L., Lobanov B. M.* (2009) Syntactic correlates of prosodically marked elements of the sentence and their role in the tasks of text-to-speech synthesis // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2009» [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii «Dialog 2009»].
5. *Jacobson, Roman* (1956) Die Verteilung der stimmhaften und stimmlosen Geräuschlaute in Russischen // Festschrift für Max Vasmer. Berlin.
6. *Khomitsevich O. G., Solomennik M. V.* (2010) Automatic pause placement in a Russian text-to-speech system // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2010» [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii «Dialog 2010»].
7. *Kibrik A. A., Kodzasov S. V., Khudyakova V. V.* (2009) Prosodic transcription: levels of detail // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2009» [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii «Dialog 2009»].
8. *Knyazev S. V.* (1999), On the perservative assimilation in Russian [O progressivnoj assimilyatsii v russkom yazyke] // Vestnik MGU, ser. 9, Philology. № 4.
9. *Knyazev S. V.* (2010), On the exceptional palatalization [O myagkosti neobychajnoj] // Phonetics and grammar: present, past, future [Fonetika i grammatika: nastoyashchee, proshedshee, budushchee] — Vol. XIII from Issues in Russian linguistics [Voprosy russkogo yazykoznaniya]. Moscow, MGU.
10. *Knyazev S. V., I. I. Vorontsova & I. V. Petrova*. (2007) Voice Coarticulation Across Word Boundaries in /v/+/v/ Sequences in Standard Russian // AFCP Workshop «Coarticulation : cues, direction, and representation». Montpellier, France. December 7, 2007 (Workshop de l'association Francophone de la Communication Parlée La co-articulation: Indices, Direction et Représentation. Organisé conjointement avec les laboratories/ DIPRALANG (EA 739) et PRAXILING (UMR CNRS 5267).
11. *Korotaev N. A.* (2009) Corpus study of pausation at syntactic borders: why pauses don't always appear where we expect them? // // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2009» [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii «Dialog 2009»].
12. *Krivnova O. F.* (1995) Perception and semantic relevance of prosodic breaks in spoken text [Pertseptivnaja i smyslovaja znachimost prosodicheskih shvov v svjaznom tekste] // Problems of Phonetics, V. 2, Science, Moscow, pp. 229–238.
13. *Krivnova O. F.* (1999) Semantic significance of prosodic breaks in spoken text [Smyslovaja znachimost prosodicheskih shvov v svjaznom tekste] // Problems of Phonetics, V. III, Science, Moscow, pp. 247–257.
14. *Krivnova O. F.* (2015) The depth of prosodic breaks in spoken text (experimental data) // Computational Linguistics and Intellectual Technologies: Proceedings

of the International Conference «Dialogue 2015» [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii «Dialog 2015»], RGGU.

15. *Lobanov B. M., Getsevich Iu. S.* (2011) Statistical characteristics of syntagmatic segmentation of utterances from the viewpoint of expressive text-to-speech synthesis // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2011» [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii «Dialog 2011»], RGGU.
16. *Panov M. V.* (1979) Modern Russian: Phonetics [Sovremennyj russkij yazyk: Fonetika], Education, M.
17. *Paufoshima R. F.* (1969) Some problems of voice/voiceless consonants category in Russian dialects [Nekotorye voprosy, svyazannye s kategoriej glukhosti/zvonkosti soglasnykh v govorakh russkogo yazyka] // Experimental phonetic study of Russian dialects [Eksperimental'no-foneticheskoe izuchenie russkikh govorov]. Nauka, M.
18. *Potapova R. K., Blokhina L. P.* (1986) Ways of prosodic phrasing of speech in German and Russian [Sredstva foneticheskogo chleneniya rechevogo potoka v nemetskom I russkom yazykakh]. Moscow.
19. *Svetozarova N. D.* (1982) The system of Russian intonation [Intonacionnaya sistema russkogo yazyka]. Leningrad.
20. *Svetozarova N. D.* (2000) The role of phrase prosody in speech and its' modelling [Rol' frazovoj intonatsii v rechevoj deyatel'nosti i vozmozhnosti yeyo modelirovaniya] // Phonology of speech [Fonologiya rechevoj deyatel'nosti]. SPbGU.
21. *Shcherba L. V.* (1915) Eastern Sorbian language [Vostochnoluzhitskoye narechie] Petrograd.
22. *Tsirulnik L. I., Lobanov B. M., Sizonov O. G.* (2008) Algorithm of the intonation marking of narrative senteces for tts synthesis // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2008» [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii «Dialog 2008»].

«КАК БЫ НЕ Я И КАК БЫ НЕ С ТОБОЙ»: ПРАГМАТИКА РЕФЕРЕНЦИАЛЬНОГО СМЕЩЕНИЯ В УСТНОЙ РЕЧИ

Колмогорова А. В. (nastiakol@mail.ru)

Сибирский федеральный университет, Красноярск, Россия

Ключевые слова: коммуникативное лицедейство, референциальное смещение, коммуникация, иллокутивная сила, прагматическая функция

"AS IF IT WASN'T ME AND IF IT WASN'T YOU": PRAGMATICS OF REFERENTIAL GLIDE IN SPOKEN LANGUAGE

Kolmogorova A. V. (nastiakol@mail.ru)

Siberian Federal University, Krasnoyarsk, Russia

The article explores pragmatic functions of the "communicative mummery"—phenomenon observable in everyday communication and consisting in some referential slides that happen in the communicative act trivial schema "I talk to You Here and Now": the speaker can communicate as if he wasn't himself but someone else or if he was talking to another person but not to his real communicative partner. Unlike a simple trickery the communicative mummery isn't hidden from the speaker's interlocutor. Conversely, experiencing such transfigurations (I speak, if it wasn't me…etc.) the speaker intentionally uses a range of prosodic and/or non verbal markers such as special gaze, gesturing that isn't familiar to him, specific accent or prosodic contours for attracting his partner attention. The discursive manifestations of communicative mummery have some common features with the reported speech and the polyphonic conversational humor phenomena but, in the same time, display its own particulars properties and perform rather special functions in conversation. Firstly remarked in mother-child communication as a particular mother's practice of child socialization the discussed phenomenon was also found in adults' heterogeneous speech interactions that, after having been collected in a corpus of 52 items, served as a data for our analysis. It shows that the main pragmatic functions of communicative mummery is to prevent the speaker's social face loss in case if he violates social conventions regulating communicative behavior in the situations of social enforcement, social guilt or self-praising.

Key words: communicative mummery, referential slides, communication, illocutionary force, pragmatic function

Введение

Суть анализируемого в данной публикации феномена состоит в том, что говорящий в какой-то момент своей устной диалогической интеракции, используя видимое его собеседнику референциальное смещение, начинает говорить и вести себя как кто-то другой или он сам, но как будто бы не здесь и не сейчас, или он сам, но не в том эмоциональном состоянии, в котором он действительно сейчас находится, или он говорит как бы не с тем человеком, который, действительно, является его собеседником. Этот процесс напоминает процесс актёрского перевоплощения, но, в отличие от последнего, происходит на глазах «зрителя» и на короткое время, не подлежит эстетической оценке, но призван реализовать иллокутивные интенции «актёра», который не обладает никакими профессиональными навыками перевоплощения. Суть такого коммуникативного явления сводится к моделированию какой-либо социально рискованной интеракции с точки зрения различных её субъектов. В итоге такая «полифония» позволяет снизить угрозу для социального лица говорящего.

1. Критерии отграничения коммуникативного лицедейства от других форм взаимодействия своего и чужого в речи

Проблема «своего» и «чужого» активно разрабатывается в лингвистике [Баранов 1994; Падучева 2011; Плунгян 2008]. Исследуются семантика и прагматика так называемых «ксено-показателей» [Арутюнова 1990] — частиц *де*, *мол*, *дескать* как знаков чужого голоса, отчуждаемой речи, чужого мира в дискурсе.

Д. Вайсом по аналогии с ксено-показателями на материале речевого пространства Государственной Думы было введено понятие ксенотекста, понимаемого как прецедентный текст и/или простые малоизвестные цитаты из чужой, внепарламентской речевой среды, используемые депутатами Государственной Думы во время парламентских дебатов [Вайс 2012; 2014]. Ксенотексты, как отмечает Д. Вайс, вводятся в речь либо особыми метатекстовыми «скрепами» (*как говорится*), либо именованием автора (*как у известного литературного героя; как сказал поэт*) или номинацией жанрового маркера цитируемого произведения (*так вот притча такая; я расскажу новогодний анекдот*) [Вайс 2014: 44].

В рамках традиции конверсационного анализа Б. Приего-Вальверде и Р. Бертран [Bertrand, Priego-Valverde 2011] вводят термин «конверсационный полифонический юмор» для обозначения случаев введения чужого голоса в диалогическом неформальном общении для характеризации с юмористическим оттенком какого-либо персонажа-предмета беседы, т. е. в тех случаях, когда двое говорящих обсуждают кого-то третьего и один из них или оба последовательно прибегают к включению реплик, которые могли бы сказать/ сказали «другие» об этом человеке.

В свою очередь, мы предлагаем использовать термин «коммуникативное лицедейство» (далее — КЛ) для описания несколько другой группы случаев.

Речь идёт о ситуации диалогического общения, не связанного с ситуацией дескрипции для получения юмористического эффекта, но — с непосредственной интеракцией, преследующей иллокутивные цели убеждения, оказания влияния, воздействия в конфликтной и/или угрожающей социальному лицу ситуации. При этом в одном и/или нескольких «узлах» коммуникативной схемы «Я ГОВОРЮ ЗДЕСЬ и СЕЙЧАС с ТОБОЙ» происходит референциальный сдвиг, который можно эксплицировать через коннектор «как бы не»: «как бы не Я / ГОВОРЮ/ как бы не ЗДЕСЬ и как бы не СЕЙЧАС как бы не с ТОБОЙ». В отличие от «ксенотекстов», в такого рода — «лицедейских» — дискурсивных фрагментах нет «метатекстовых скреп». Говорить о «чужом слове» здесь можно лишь в случае наличия референциального сдвига «как бы не я», но и в подобных примерах отсутствуют ксено-показатели. В остальных же случаях следует, скорее, говорить о появлении особой модальности «понарошку» как некой коммуникативной игры, проявляющейся именно в ситуациях угрозы коммуникативному лицу говорящего. Маркерами феномена, который мы называем «коммуникативным лицедейством» являются:

а) референциальный сдвиг в узлах коммуникативной схемы, манифестируемый через «незаконное» смещение дейктических указаний [Падучева 2011] (*я* для обозначения другого субъекта — не реального говорящего, *он*, *она* — для *ты)*;

в) необычная для говорящего просодия (резкая смена тембра голоса, появление несвойственного акцента, резкое изменение высоты и силы голоса, резкая смена интонационного рисунка);

с) изменения на уровне невербального поведения (появление несвойственных говорящему и/ или не предполагаемых ситуацией жестов, движений, иногда — миметического характера).

2. Корпус

Первоначально обсуждаемый феномен был отмечен нами при работе с 90 часовым видеокорпусом записей общения матерей со своими детьми в возрасте от 0 до 7 лет. В общей сложности было проанализировано 120 диад. Феномен, названный нами «коммуникативным лицедейством» регулярно наблюдался в речи матерей, обращённой к детям от 0 до 3-х лет, затем — появлялся лишь окказионально, а при общении с детьми 6–7 лет вообще не был отмечен. Таким образом, КЛ наблюдалось в псевдоконфликтных (моделируемых как конфликтные) коммуникативных ситуациях: 1) в присутствии третьих лиц (чаще — родственниц) мать, обращаясь к ещё неспособному говорить ребёнку, произносит обвинительные реплики в свой адрес от имени ребёнка (*мать, скажи, не кормит меня*… и т.д.); 2) в присутствии ребёнка моделирует от его имени реплики возмущения, обращённые к третьему лицу (*скажи, мне не нравится, дядя, мне не нравится*); 3) говорит о себе в третьем лице в ситуации недозволенного, неодобряемого социумом поведения ребёнка (*мама обиделась и заплакала, пожалей маму*).

Как результат работы с данным корпусом появилась гипотеза о том, что референциальное смещение подобного типа может использоваться и в устной речи взрослых, но для какой цели?

Дальнейший сбор материала для «взрослого» корпуса проводился методом случайной выборки, отсюда — гетерогенный характер материала (фрагменты художественных фильмов, опосредованных художественным дискурсом диалогов, интервью известных медийных личностей, записи повседневной устной речи, сделанные автором, общее количество единиц «взрослого» корпуса — 52 дискурсивных фрагмента), обусловленный сравнительной редкостью изучаемого феномена в речи взрослых, а также его спонтанностью: он носит мультимодальный характер, возникает в диалогическом общении, как правило, повседневном и неформальном или приближенном к таковому.

3. Коммуникативное лицедейство, тип 1: референциальное смещение «как бы не я»

Рассмотрим способы маркирования в речевом поведении и прагматические функции различных типов коммуникативного лицедейства, отталкиваясь от критерия «тип референциального смещения». Первый тип — референциальное смещение «как бы не я», в котором, собственно, и слышен «чужой голос».

В фильме «Ликвидация» (реж. С. Урсуляк) друг, и «по совместительству» — вор-карманник, Фима уговаривает Давида Гоцмана заглянуть на минутку к друзьям, но Давид отказывается. Тогда Фима внезапно переходит от типичной для него одесской речи к специфической интонации и немецкому акценту:

(1) *Г.: /Фима/ мне надо до военных прокуроров/ а ты вцепился что лешай до пионерки/ поясни хоть/по какому случаю//*
Ф.: /три минуты и побежишь до своих прокуроров/ три минуты/ ***друг*** *просит//*
Г.: /Фима..↑/
Ф.: # резко останавливается, одной рукой быстро резко берёт Гоцмана за плечо, другой — как бы вытаскивает их кармана пистолет, складывает средний и указательный пальцы в виде дула, которое направляет в грудь собеседнику, смотрит прямо в газа#
нjэт↑ ви ***<u>пойдйоте</u>*** *со мной генерал'↓/*
Г.: #ухмыляется и идёт за Фимой#

В данной ситуации говорящий на очень короткое время нарушает референциальную схему коммуникативной ситуации, говоря, как если бы это был не он, а немецкий офицер, следовательно, как если бы действие происходило не в послевоенной Одессе, а во время оккупации где-нибудь в концлагере или лагере военнопленных — референциальное смещение в узле «субъект коммуникации», как правило, ведёт к смещению и в других узлах схемы коммуникативного акта.

Своего рода начальным маркером коммуникативного лицедейства в данном случае выступают, прежде всего, невербальные средства: проксемические — Фима резко останавливается, жестовые — резко берёт собеседника за плечо. При дальнейшем развёртывании «перевоплощения» линия несвойственного говорящему невербального поведения продолжается: резкие властные, сдержанно агрессивные жесты, взгляд, демонстрируемые Фимой, не имеют ничего общего с его обычным поведением до и после анализируемого фрагмента. Отметим также имитацию воображаемого действия «вытаскивать пистолет» как невербальный маркер коммуникативного лицедейства.

Просодически хорошо заметно появление характерного немецкого акцента, реплика произносится более напряжённым тоном и громче, а также быстрее, чем предыдущие и последующие, характерна интонация выраженной иллокутивной силы [Арутюнова 1999: 672], характерная для институционального, а не межличностного общения.

Прагматическая функция появления в данной части интеракции чужого голоса тесно связана с ведущей иллокуцией в данном дискурсивном фрагменте — директивой или (по Дж. Личу [Leech 1983]) конкурирующей — и состоит в решении нескольких вспомогательных задач: 1) с помощью переключения с фамильярного регистра речи на институциональный усилить в ситуации **социального принуждения** императивность иллокуции до уровня, который в межличностном дружеском общении привёл бы к риску прерывания коммуникации; 2) при этом, одновременно с усилением императивности, уменьшить степень угрозы собственному социальному лицу в контексте отношений дружбы, приятельства, «разрядив обстановку» за счёт видимого эффекта перевоплощения. «Чужой голос» здесь призван показать реальную иллокутивную силу, заложенную в высказывании говорящим, выраженную в недопустимой в реальной социо-коммуникативной ситуации форме.

Подобный тип коммуникативного лицедейства встречаем и в институциональном общении, приближенном, однако, к повседневному в риторических и имиджевых целях. Так, В. В. Жириновский, выступая во время пресс-конференции накануне президентских выборов 2011 в качестве одного из кандидатов, использует подобный способ введения «чужих голосов» с референциальным смещением. Особенностью введения «лицедейского фрагмента» является его лексическое анонсирование, даже создание своего рода иконических отношений между семантикой некоторых лексем и просодикой следующего за ними контекста. Например, произносит, говоря о В. В. Путине, «ему ближе тихие чиновники» тихим бесстрастным голосом, а сразу после «он не понимает, что страна требует окрика» кричит «Молчать!». Завершение же «лицедейства» выделяется метатекстовым комментарием («вот это страна почувствует, вот это царь!»):

(2) */ я его раздражаю/ ему ближе тихие чиновники*…..[0,2] *кабинетные #говорит тихим монотонным голосом# он сидит и тихо разговаривает/ он не понимает/ что страна требует окрика//*
***МОЛЧАТЬ!** #кричит, бьёт кулаком по столу#*
/вот это страна почувствует/ вот это царь! император! ***ВСТААТЬ!***

#подбородок приподнят, жест «указующий перст», высокомерный взгляд на присутствующих «сверху-вниз»
#тихо, с элементами передразнивания# *садитесь/ ну как у вас там дела/зарплату/ ну/ надо повысить/ учителям/ ну среднеэкономическая по стране/ вот у вас ЖКХ вроде бы повысилось* [*разводит руками*] [*кричит*] **БЕЛЫХ, Я ТЕБЯ ВЫГОНЯЮ ! ТЫ ТАМ В ШВЕЦИИ, ТАК И ОСТАВАЙСЯ В ШВЕЦИИ** [*жест «указующий перст»*] **У ТЕБЯ ТАМ ПОВЫШЕНИЕ 40% ЗА ГАЗ И ЗА СВЕТ**//

По-видимому, в анализируемом фрагменте к собственно коммуникативному лицедейству можно отнести лишь те части, где «Я» говорящего локализовано в образе царя, поскольку речь В. В. Путина, скорее, передаётся в модусе пересказывательности [Левонтина 2010]. Свидетельства тому — наличие маркера «передачи угрозы и осуждения в чужой речи» [Подлесская, Кибрик 2009] *вот*, а также *ну*, которое маркирует пересказ чужой речи, оцениваемой как не содержащая ничего нового, повторяющая давно известные вещи.

В части «царского голоса» сохраняются и выделенные нами ранее типичные маркеры коммуникативного лицедейства: референциальное смещение типа «как бы не я», влекущее за собой и другие смещения («как бы не здесь» и «как бы не сейчас») — Жириновский коммуницирует как «как бы» царь (видимо, Иван Грозный); отсутствуют прямые маркеры говорения, которые вводили бы «чужой голос», но резкая смена просодики и невербального поведения удивляет неожиданными перепадами; появляются элементы миметических схем — имитирует удары посохом (скипетром).

При этом подчеркнём, что в случаях, описываемых как «коммуникативное лицедейство», при том, что происходит «разрыв» личностной самоидентификации, чужой голос ассоциируется говорящим с собой как субъектом коммуникации, поэтому такой голос не передаётся ни в аспекте пересказывательности с различными модальными оттенками, ни в аспекте передразнивания или утрирования.

Прагматические функции КЛ в проанализированном дискурсе В. Жириновского: 1) усилить имиджевую составляющую, показав неконвенциональный, но эффективный способ решения насущных проблем; 2) при этом, одновременно, уменьшить степень угрозы собственному социальному лицу в глазах политической элиты, поскольку подобная коммуникация недопустима в контексте институциональных отношений.

Частым видом референциального смещения «как бы не я», относимого нами также к КЛ, является разговор о себе в третьем лице — так, как если бы о говорящем высказывался некто другой:

(3) *Вера Ивановна: А/ вы чего сидите/ ждёте/когда Верывановна* ***придёт*** *и* ***всё решит****/ нет давайте сами//*
/можт всё-таки помооожете/ Верывановна↑//

В примерах данного, последнего, типа используя «он-перспективу», говорящий помещает себя в позицию агенса, а выполняемые им действия попадают

в фокус коммуникативного внимания. Прагматика подобных высказываний направлена на социальное повышение говорящего, являясь формой скрытой лести самому себе. Однако за счёт референциального смещения говорящий не нарушает социальных конвенций и не наносит ущерба своему социальному лицу: он подчёркивает свою важность и незаменимость, говоря «от имени и по поручению» третьих лиц.

4. Коммуникативное лицедейство, тип 2: референциальное смещение «как бы не с тобой»

Примером второго рода считаем коммуникативную ситуацию в фильме «Такси — 3» (русскоязычная версия), где главный герой (Д.) объясняется со своей девушкой (Л.): она ушла от него, поскольку считала, что он не уделяет должного внимания отношениям и посему будет плохим отцом. Для обозначения своей собеседницы он выбирает существительное, эквивалентное местоимению 3 лица единственного числа, осуществляя таким образом референциальное смещение «я говорю здесь и сейчас как бы не с тобой». Девушка (Л.) подхватывает эту игру:

(4) *Л.: /**ты** тут зачем↓/ ты **что**/ беременный↑/*
*Д.: / #*медленно и расслабленно, слегка театрально, начинает приближаться к девушке, выразительно смотрит*# моя **любимая** беременна↓/*
Л.: /твоя любимая забеременела от выхлопной трубы ↑или от сквозняка ↓/
Д.: /от меня/ [0,1] *я пообещал любимой*
*Л.: /**что**↑ пообещал↓/*
Д.: /что буду больше времени проводить с ней/ чем с машиной/ что буду лучшим на свете папой…
Л.: /и она поверила↑/
Д.: /не↑ уверен↓/#улыбается # но это будет самой большой ошибкой в её жизни/ потому что я говорю это искренне/
Л.: / у вас искренности в изобилии/ а времени нет↑/ чтобы выполнить все обещания↓/
Д.: /Лили↑/ ты выйдешь за меня↑/

В анализируемом дискурсивном фрагменте референциальное смещение «Я /говорю/ как бы не с тобой» инициировано одним коммуникантом и подхвачено другим, что привело к референциальному смещению и в речи второго коммуниканта «как бы не Я». Далее мы остановимся, тем не менее, на особенностях коммуникативного поведения первого из них.

По-видимому, к этому же типу относится записанный нами разговор в институциональной «рабочей» ситуации, где первый говорящий (Г1) занимает более высокую социальную позицию и обращается к собеседнику как если бы говорил не с ним, но о нём:

(5) *Г1: /Я уже не раз говорила* ***одному коллеге***
#выразительный взгляд# /
что не стоит так вести себя в коллективе/ не припомните/ Андрей Петрович/говорила↑//
Г2: / Говорили/ помню/ ну и как↑ ***коллега*** *↓[0,01] исправился ↑//*
Г1: /По-видимому ***нет****↓/*
*Г2: /Ну/ ре****дии****ска\/#выразительный взгляд#*
Г1: #молчит, выразительно смотрит#
Г2: /Ну всё↓/ больше не буду↓/ простите↓//

В примерах 4–5 референциальное смещение «я говорю как бы не с тобой» маркируется в самом начале «лицедейства» акцентным выделением либо слова, имеющего в качестве референта реального собеседника говорящего (моя **любимая** беременна), либо определения, нарочито подчёркивающего неопределённость референта (**одному** коллеге), сопровождаемым и невербальным маркером — выразительным взглядом. В обоих случаях собеседники подхватывают «игру в лицедейство», которая для демонстрации перлокутивного эффекта в конечной фазе интеракции обрывается тем из них, кто чувствует за собой вину. Он ликвидирует референциальное смещение, показывая своё понимание того, кто является реальным адресатом (в первом фрагменте — это просящий извинения инициатор «лицедейства», во втором — также извиняющийся коммуникативный партнёр инициатора «лицедейства»). Кроме того, в обоих проанализированных случаях заметен специфический просодический характер высказываний инициатора КЛ: слегка замедленный, но не утрированно, темп речи, растягивание гласных, повторяющийся плавный восходященисходяций интонационный контур.

Что касается прагматических функций КЛ в проанализированных дискурсивных фрагментах, то отметим следущее: по-видимому, разговор «как бы не с тобой (но о тебе)» позволяет инициатору КЛ в ситуации «**социальной вины**» увеличить силу иллокуции до уровня, который в обычной коммуникации угрожал бы его социальному лицу (например, истово умолять о прощении (пр.4) или угрожать и осуждать виновника (пр. 5)), избежав при этом коммуникативных рисков за счёт удвоения позиции своего партнёра по коммуникации — он, понимая, что иллокутивная сила направлена на него, одновременно выступает и в позиции наблюдателя со стороны.

5. Коммуникативное лицедейство, тип 3: референциальное смещение «как бы не здесь и не сейчас»

Следующий из рассматриваемых видов КЛ мы многократно наблюдали в видеозаписях общения матерей со своими детьми в возрасте от года до 2-х-3-х лет. В ситуациях, рассматриваемых матерями как достойные порицания и гипотетически способные попасть в поле речевой агрессии, они коммуницировали с ребёнком так, как будто бы были агрессивны, разозлены, однако

отдельные маркеры выдавали «понарошечный» харакер этого состояния, т. е. матери вели себя так, как они бы вели себя в другой ситуации («не здесь и не сейчас»), если бы были действительно разозлены и агрессивны :

(6) *Мама: /Почему не спишь/ зараза такая/* ***почему*** *не спишь/ а↑/ ты чо лежишь там пищишь↑/ну-ка спи# быстро#/ложись давай* [0,02]*/ ну-ка* ***спи*** *говорю↓/комугрю# быстро# комугрю спи ложись/ я тебе/ ой я тебе/ Марусина/ Ма\/русина//*

Основные маркеры такого КЛ — невербальные и просодические. Среди первых отметим некую неполноту амплитуды агрессивных действий: если машет кулаком, то делает это «мелко-мелко» потрясывая, если грозит пальцем, то тоже быстро и короткими движениями; смеющийся взгляд. Среди просодических маркеров выделим неестественно быстрый темп речи, её неестественно высокий тон там, где в «обычной» агрессии наблюдались бы его падения (например, в сегментах, передающих императивную коммуникативную тональность), общий сдвиг вперёд артикуляции.

Реализации подобных коммуникативных моделей, имеющих, в отличие от материнского общения, выраженную прагматическую составляющую, находим в бытовом и институциональном общении взрослых. Ср.:

(7) Н. (начальник)*: /Ну* [0,01] *идите сюда/* ***есть*** *серьёзный разговор#хмуро перекладывает бумаги на столе #/ #смотрит в упор #Вот вам* ***сколько*** *лет↑*
П.: /#напряжённо#сорок/
Н.: /и вот ***скОлько*** *можно повторять* [0,01] *# выразительно смотрит#вдруг улыбается# с днём* ***рождЕния, с сорокалетием****↓/*

При анализе видеозаписи (пр. 7) отмечаем, хотя и в ослабленном виде, те же, что в пр. 6 маркеры: слегка ускоренный темп речи инициатора КЛ, общее «завышение» тона в высказывании по сравнению с «обычной» агрессивной коммуникацией. Среди невербальных маркеров — отсутствие агрессии во взгляде, сдержанный характер жестикуляции. Хотя на первый взгляд в случае пр. 7 мы можем квалифицировать наблюдаемое как шутку, однако трудно согласиться с тем, что КЛ здесь имеет исключительно развлекательный эффект — в ситуации социальной зависимости осуществляется как бы «проверка» и укрепление таких неравных социальных отношений, не приводящая при этом к потере инициатором КЛ социального лица (например, если бы директор (пр. 7) спросил «Вы боитесь меня, вы понимаете, что я ваш начальник?»).

Таблица 1. Маркеры и прагматические функции референциального смещения в устной речи

Говорю…	Вербально-просодические маркеры	Невербальные маркеры	Частно-прагматическая функция	Общая прагматическая функция	Кол-во в корпусе
«как бы не я»: Я-перспектива	несвойственные говорящему акцент, интонация, резкая смена ритма	несвойственные говорящему жесты и мимика, имитационные жесты	неконвенциональное усиление императивности	Снижение угрозы сосоциальному лицу +	14
Он-перспектива	акцентное выделение предикатов действия	движение головой сверху вниз с выдвижением вперёд подбородка	положительная самопрезентация, самовосхваление		17
«как бы не с тобой»	акцентное выделение лексем-номинаций собеседника, замедленный темп, растягивание гласных	выразительный взгляд на собеседника	увеличение иллокутивной силы высказывания в ситуации социальной вины		12
«как бы не здесь и не сейчас»	ускоренный темп, завышение тона	отсутствие агрессии во взгляде и жестах	проверка вертикальных социальных отношений		9

Заключение

Прагматической функцией, общей для всех рассмотренных случаев КЛ, является предотвращение потери говорящим — инициатором КЛ — собственного социального лица при одновременном нарушении им конвенций одобряемого социумом в данной ситуации типа коммуникативного поведения.

Так, КЛ с референциальным смещением «как бы не я», по всей видимости, свойственно ситуациям социального принуждения, когда говорящий прибегает к таким его коммуникативным формам реализации, которые не одобряются социумом (я-перспектива). При этом просодические и невербальные маркеры КЛ ярко представлены. В случае он-перспективы («как бы не я, а кто-то обо мне») прагматическая функция — самовосхваление.

КЛ с референциальным смещением «как бы не с тобой» реализуется в ситуациях, связанных с социальной виной инициатора КЛ или его собеседника. При этом ведущую роль играют просодические маркеры, в том числе — акцентное

выделение первого существительного, номинирующего реального собеседника в он-перспективе.

КЛ с референциальным смещением «как бы не здесь и не сейчас» наблюдается в ситуациях социальной зависимости, когда занимающий более высокую позицию на социальных качелях инициатор КЛ хочет проверить такие вертикальные социальные связи «на крепость», подтверждая тем самым свою позицию в иерархии. В данном случае ярко выраженными оказываются просодические и невербальные маркеры КЛ.

Литература

1. *Арутюнова Н. Д.* Высказывание в контексте диалога и чужой речи // Revue des études slaves, Tome 62, fascicule 1–2, 1990. L'énonciation dans les langues slaves [En hommage à René L'Hermitte, sous la direction de Jean-Paul Sémon et Hélène Włodarczyk]. Pp. 15–30.
2. *Арутюнова Н. Д.* Чужая речь: «своё» и «чужое» // Язык и мир человека. М.: Языки русской культуры, 1999. С. 668–686.
3. *Баранов А. Н.* Заметки о дескать и мол // Вопросы языкознания. 1994. № 4. С. 114–124.
4. *Вайс Д.* Депутаты любят цитаты: ссылки на ксенотексты в Госдуме // Русский язык сегодня. Вып. 5: Проблемы речевого общения. М.: ФЛИНТА: Наука, 2012. С. 64–75.
5. *Вайс Д.* Интертекстуальность в Госдуме //Политическая коммуникация: перспективы развития научного направления: материалы Междунар. науч. конф. (Екатеринбург, 26–28.08.2014) / гл. ред. А. П. Чудинов. Екатеринбург, 2014. С. 43–45.
6. *Левонтина И. Б.* Пересказывательность в русском языке // Компьютерная лингвистика и интеллектуальные технологии. Вып. 9 (16). По материалам международной конференции Диалог 2010. С. 284–288.
7. *Николаева Т. М.* От звука к тексту. М.: Языки русской культуры, 2000. 679 с.
8. *Падучева Е. В.* Семантические исследования. Семантика времени и вида в русском языке; Семантика нарратива; изд.2-е испр. и доп. М.:Языки русской культуры, 2011. 480 с.
9. *Падучева Е. В.* Показатели чужой речи *мол* и *дескать* // Известия РАН. Серия литературы и языка. 2011. Т. 70, N 3. С. 13–19.
10. *Плунгян В. А.* О показателях чужой речи и недостоверности в русском языке: мол, якобы и другие // B. Wiemer & V. A. Plungjan (Hrsg.). Lexikalische Evidenzialitäts-Marker in slavischen Sprachen // Wiener Slawistischer Almanach, Sonderband 72. München: Sagner, 2008. S. 285–311.
11. *Подлесская В. И., Кибрик А. А.* Дискурсивные маркеры в структуре устного рассказа: опыт корпусного исследования // Диалог 2009. С. 390–396.
12. *Bertrand R., Priego-Valverde B.* Does prosody play a specific role in conversational humor? Pragmatics and Cognition, vol. 19, no. 2. 2011, p. 333–356.
13. *Leech G.* Principles of Pragmatics. London, New York: Longman, 1983. 257 p.

References

1. *Arutjunova N. D.* (1990), Utterance in the context of dialogical and reported speech [Vyskazyvanie v kontekste dialoga i chuzhoj rechi], Journal of Slavistic Studies [Revue des études slaves], Tome 62, fascicule 1–2, 1990. Utterance Organization in Slave Languages [L'énonciation dans les langues slaves] [En hommage à René L'Hermitte, sous la direction de Jean-Paul Sémon et Hélène Włodarczyk]. pp. 15–30.
2. *Arutjunova N. D.* (1999), Reported speech: "my own" and "others'" [Chuzhaja rech': «svojo» i «chuzhoe»], Language and Human World [Jazyk i mir cheloveka], Jazyki russkoj kul'tury, Moscow, pp. 668–686.
3. *Baranov A. N.* (1994), Notes about *deskat'* and *mol* [Zametki o deskat' i mol], Issues of General Linguistics [Voprosy jazykoznanija], Vol.4, pp.114–124.
4. *Bertrand R., Priego-Valverde B.* (2011), Does prosody play a specific role in conversational humor? Pragmatics and Cognition, vol. 19, no. 2. 2011, pp. 333–356.
5. *Leech G.* (1983), Principles of Pragmatics. London, New York: Longman.
6. *Levontina I. B.* (2010), Reported speech in Russian [Pereskazyvatel'nost' v russkom jazyke], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2010" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2010"], Bekasovo, pp. 284–288.
7. *Nikolajeva T. M.* (2000), From the sound to the text [Ot zvuka k tekstu], Jazyki russkoj kul'tury, Moscow.
8. *Paducheva E. V.* (2011), Markers of reported speech *mol* and *deskat'* [Pokazateli chuzhoj rechi mol i deskat'], Russian Acedemy of Sciences Bulletin, Litterature and Language2011. Vol. 70, N 3. C. 13–19.
9. *Paducheva E. V.* (2011), Semantic studies. Semantics of tense and voice in Russian. Semantics of Narrative [Semanticheskie issledovanija. Semantika vremeni i vida v russkom jazyke; Semantika narrativ], Jazyki russkoj kul'tury, Moscow.
10. *Plungjan V. A.* (2008), Remarks on the reported speech markers and non-reliability in Russian: mol, jakoby and others [O pokazateljah chuzhoj rechi i nedostovernosti v russkom jazyke: mol, jakoby i drugie], Lexical Markers of Evidence in Slave Speech [Lexikalische Evidenzialitäts-Marker in slavischen Sprachen], Slavistic Almanac of Vienne, Vol. 72. München: Sagner, pp. 285–311.
11. *Podlesskaja V. I., Kibrik A. A.* (2009), Discursive markers in the structure of oral narration: the case of corpus study [Diskursivnye markery v strukture ustnogo rasskaza: opyt korpusnogo issledovanija], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2009" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2009"], Bekasovo, pp. 390–396.
12. *Weiss D.* (2012), Deputies love quatations: references to xeno-texts in Duma [Deputaty ljubjat citaty: ssylki na ksenoteksty v Gosdume], Russian Language Today [Russkij jazyk segodnja], Vol. 5: Problems of Verbal Communication, Flinta-Nauka, Moscow, pp. 64–75.

13. *Weiss D.* (2014), Intertextuality in Gosduma [Intertekstual'nost' v Gosdume], Political Communication: Perspectives of Research Approach: Procedings of the International Conference [Politicheskaja kommunikacija: perspektivy razvitija nauchnogo napravlenija: materialy Mezhdunar. nauch.konf.], Ekaterinburg, pp. 43–45.

ТОНАЛЬНЫЙ СЛОВАРЬ И ОБУЧАЮЩАЯ КОЛЛЕКЦИЯ ДЛЯ СЕНТИМЕНТ-АНАЛИЗА СОЦИАЛЬНО-ПОЛИТИЧЕСКИХ ТЕКСТОВ

Кольцова О. Ю. (ekoltsova@hse.ru)[1],
Алексеева С. В. (salexeeva@hse.ru)[1, 2],
Кольцов С. Н. (skoltsov@hse.ru)[1]

[1]НИУ ВШЭ, Санкт-Петербург, Россия

[2]СПбГУ, Санкт-Петербург, Россия

Ключевые слова: словарь тональной лексики, веб-интерфейс, краудсорсинг тональной разметки, российская блогосфера, «Живой журнал», размеченная коллекция, тематическое моделирование

AN OPINION WORD LEXICON AND A TRAINING DATASET FOR RUSSIAN SENTIMENT ANALYSIS OF SOCIAL MEDIA

Koltsova O. Yu. (ekoltsova@hse.ru)[1],
Alexeeva S. V. (salexeeva@hse.ru)[1, 2],
Kolcov S. N. (skoltsov@hse.ru)[1]

[1]National Research Institute Higher School of Economics, St. Petersburg, Russia

[2]St. Petersburg State University, St. Petersburg, Russia

Automatic assessment of sentiment in large text corpora is an important goal in social sciences. This paper describes a methodology and the results of the development of a system for Russian language sentiment analysis that includes: a publicly available sentiment lexicon, a publicly available test collection with sentiment markup and a crowdsourcing website for such markup. The lexicon is aimed at detecting sentiment in user-generated content (blogs, social media) related to social and political issues. Its prototype was formed based on other dictionaries and on the topic modeling performed on a large collection of blog posts. Topic modeling revealed relevant (social and political) topics and as a result—relevant words for the lexicon prototype and relevant texts for the training collection. Each word was assessed by at least three volunteers in the context of three different texts where the word occurred while the texts received their sentiment scores from the same volunteers as well. Both texts and words were scored from

–2 (negative) to +2 (positive). Of 7,546 candidate words, 2,753 got non-neutral sentiment scores. The quality of the lexicon was assessed with *SentiStrength* software by comparing human text scores with the scores obtained automatically based on the created lexicon. 93% of texts were classified correctly at the error level of ±1 class, which closely matches the result of *SentiStrength* initial application to the English language tweets. Negative classes were much larger and better predicted. The lexicon and the text collection are publicly available at http://linis-crowd.org.

Key words: sentiment lexicon, web interface, crowdsourcing sentiment markup, Russian blogosphere, livejournal, test collection, topic modeling

1. Introduction

Sentiment analysis (SA) in Russia has so far been focused on polarity detection in customer reviews: this, for instance, can be clearly seen from the content of the Russian Information Retrieval Seminar (*ROMIP*) competition on SA (Chetviorkin et al, 2012; Chetviorkin, Loukachevitch, 2013; Loukachevitch et al, 2015). However, marketing professionals are not the only potential "consumers" of automatic sentiment analysis techniques. Social scientists get increasingly interested in "online public opinion" on various social and political issues or events, as well as in predicting public reaction to those events with online sentiment data. At the moment, no Russian language sentiment lexicons or machine learning instruments are publicly available (exception: Chetviorkin-Loukashevich dictionary of sentiment-bearing words with undefined polarity for consumer reviews in three domains). As a result, researchers in Russia can only rely on commercial services whose methodologies are never completely disclosed. This is often unacceptable for academic users.

This work seeks to make a first step in the development of freely available resources for the Russian language SA We develop a domain-specific sentiment lexicon and check its quality against the marked-up collection of political and social post fragments written by top bloggers at the most popular Russian blog platform *LiveJournal*. Our sentiment analysis task here is reduced to a relatively simple classification of texts into those with prevailing negative emotions and those with prevailing positive emotions, irrespectively of the object of these sentiments—that is, we do not solve a political support/oppose classification task.

The rest of the paper is organized as follows. We first take a visit on the previous literature. Next, we explain our data collection, sentiment lexicon formation, and the markup procedure. Then, we report word and text assessment results and analyze the quality of the lexicon. Finally, we close the paper with a conclusion.

2. Related work

SA can be conventionally divided into two main approaches (Pang, Lee, 2008; Medhat et al, 2014):

(1) Lexicon-based approach (Taboada et al, 2011). It browses texts for certain words or phrases whose polarity has been predefined, often in relation to the domain of interest. Such thesauri are often supplemented with a set of rules, concerning the use of negation or booster words. Some of the well-known limitations of this approach are domain sensitivity and initial lexical insufficiency while its simplicity is one of its main advantages.

(2) Machine learning approach. It uses marked-up text collections (training datasets), as well as feature lists, as information which a mathematical algorithm relies on while classifying other marked-up collections (test sets). Most of such algorithms optimize until the best possible fit with the test set markup is reached. After that, these algorithms are applied to non-marked-up (real world) collections. This more sophisticated approach most often yields better results, however, it is vulnerable for overfitting and requires large marked-up corpora to produce high quality.

In addition, these two approaches work differently for different tasks. For instance, SVM method for the task of binary classification of English-language movie reviews has yielded precision of 86.4% (Pang, Lee 2004), which is particularly high. At the same time, a lexicon-based approach has been successfully used for sentiment analysis of English language social media with *SentiStrength* system (Thelwall et al, 2010) (for more details see section 4). For the Russian language, during the *ROMIP* SA competition in 2012, the best results in consumer review classification tasks were obtained by machine learning approaches, however, in political news classification, lexicon-based approaches took the lead (Chetviorkin, Loukachevitch, 2013). The competition organizers attribute this latter success to the great diversity of topics (subdomains) occurring in the news and to the absence of a sufficient training set.

These two conditions are met by user-generated social and political content from blogs and social media, the object of our interest, which is why we have chosen the lexicon-based approach as a first step.

Two main methods of sentiment lexicon generation—manual and semi-automatic—are usually described in literature (Mohammad, Turney, 2013; Taboada et al, 2011). The manual method is a human markup of words into sentiment classes, which can be very reliable when qualified experts are used. Among limitations of this approach are its labor-intensive character (although not more intensive than in the creation of marked-up text collections) and the mentioned above initial insufficiency. The latter means that initially it is hard to think of all potentially sentiment-bearing words without additional methods of their extraction.

This problem is addressed by semi-automatic methods of lexicon generation, notably by bootstrapping techniques (Thelen, Riloff, 2002; Godbole et al, 2007). They start with small lists of words with pre-defined polarity (seed words) and automatically extend them with a number of linguistic instruments. Those include measurement of sematic

association between words (Turney, 2002), synonym/antonym dictionaries (Hu, Liu, 2004) or general dictionaries or various pre-existing taxonomies (Esuli, Sebastiani, 2005). Sentiment lexicons developed for other languages are also applied (Mihalcea et al, 2007), although in our experience their usefulness is limited. Sentiment-bearing adverbs may be automatically derived from the respective seed adjectives (Taboada et al, 2011), which is a technique we have borrowed. Chetviorkin and Loukashevitch (2012) offer a methodology of detecting sentiment-bearing words (but not their polarity) for Russian language customer reviews: having manually annotated 18,362 words, they then train a classifier to detect more sentiment-bearing words and show a good quality.

Thus, semi-automatic approaches may solve the problem of labor-intensiveness in manual lexicon construction only partially, while marked-up collections can be a solution only when they emerge without researchers' effort (e.g. consumer reviews). Classification of other types of content in resource-scarce languages faces a cold start problem. In recent years, it is increasingly often addressed with crowdsourcing, both in SA (Hong et al, 2013) and other linguistic tasks (Mohammad, Turney, 2011). Crowdsourcing, as a technique relying on cheap or free labor of a large number of lay persons, brings about its own problems, notably the issue of insufficient quality resulting from the lack of qualification or motivation. Approaches to coping with this are still in their cradle. While Hong et al (2013) suggest to motivate volunteers through gamification, Hsueh et al (2009) develop a number of methods to detect and discard low-quality assessments. The gold standard in both cases is, however, expert opinion, which itself is prone to individual biases when it comes to the polarity of political texts. In addition, resource-scarce languages may be resource-scarce precisely because crowdsourcing services are unavailable either for technical or financial reasons. We address some of these problems further below.

3. Data collection and markup

3.1. Generating relevant text collection

We extract our collection from our database that includes all posts by top 2,000 LiveJournal bloggers for the period of one year (from March 2013 to March 2014). Earlier we found out that only about a third of those texts may be classified as political or social (Koltsova et al, 2014), hence, we face a problem of retrieving relevant texts. While Hsueh et al (2009) employ manual annotation, this is unfeasible for our collection of around 1.5 million texts, so we adopt a different approach (Koltsova, Shcherbak, 2015). We perform topic modeling, namely Latent Dirichlet Allocation with Gibbs sampling (Steyvers, Griffiths, 2004). It yields results akin to fuzzy clustering, by ascribing each text to each topic out of a predefined number, with a varying probability, based on word co-occurrence. All words are also ascribed to all topics with varying probabilities. When sorted by this probability they form lists that allow fast topic interpretation and labeling by humans. Our *TopicMiner* software (http://linis.hse.ru/soft-linis) was used for all topic modeling procedures.

Our prior experience shows that the optimal number of topics depends most of all on the "size" of topics to be detected (smaller topics demand a larger number). A series of experiments (Bodrunova et al, 2013; Nikolenko et al, 2015) has lead us to choose the number of 300 for the task of retrieving social and political topics. 100 most relevant texts and 200 words of each topic were read by three annotators who have identified 104 social or political topics. The topic was considered relevant if two of the three annotators had chosen it. Intercoder agreement, as expressed by Krippendorf's alpha, is 0.578. Texts with the probability higher than 0.1 in these 104 topics (mean probability = 0.3) were considered relevant and were included into the final working collection which comprised 70,710 posts.

3.2. Selection of potentially sentiment-bearing words

Based on the aforementioned literature, we employed a complex approach to the generation of a proto-lexicon for manual annotation (all details on this approach can be found in Alexeeva et al, 2015) comprising the following elements:

- the list of high-frequency adjectives created by the *Digital Society Lab* (http://digsolab.ru) based on a large collection of Russian language texts from social media, and the list of adverbs automatically derived from the former list;
- Chetviorkin-Loukashevitch lexicon (Chetviorkin, Loukachevitch, 2012) (most of it later discarded);
- Explanatory Dictionary of the Russian Language (Morkovkin, 2003);
- Translation of the free English-language lexicon accompanying *SentiStrength* software (Thelwall et al, 2010);
- 200 most probable words for each of the relevant topics identified by annotators, which was aimed at detecting domain-specific words.

We formed a lexicon of potentially sentiment-bearing words accepting only those that occurred in at least two of the listed sources. The lexicon comprised 9,539 units. However, only 7,546 of them occurred in the texts identified as social or political, and only they were later manually annotated.

3.3. Data markup and evaluation of crowdsourcing results

To avoid some pitfalls of crowdsourcing we have adopted, so to say, a sociological vision of it: our volunteers were not supposed to imitate experts; rather, their contribution was seen as similar to that of respondents in an opinion poll which cannot produce "wrong" answers. For that, we tried to make our sample of assessors as diverse as possible in terms of region, gender, and education. In total, 87 people from 16 cities took part in the assessment.

They worked with our website linis-crowd.org and assessed words' sentiment as expressed in the texts in which they occurred, as well as the prevailing sentiment of the texts themselves, with a five-point scale, from -2 (strong negative), to +2 (strong positive). The texts were to help detect domain-specific word polarity.

Each word was shown with three different texts, one at a time. Each post was cut down to one paragraph since long texts are more likely to include different sentiments. Once each word received three annotations, we went on with further annotation to get more than one assessment for each text. By the time of data analysis, we received 32,437 word annotations and the same number of text annotations (of them, 14 word annotations and 18 text annotations were discarded due to technical errors). In total, the assessors annotated 19,831 texts. Annotated word and text collections are available at http://linis-crowd.org/.

Intercoder agreement in word assessment task (five-class), as expressed by Krippendorf's α, has turned out to be 0.541. To compare, Hong et al (2013) report α as low as 0.11–0.19 for a three-class word sentiment annotation task. Taboada et al (2011: 289) obtain mean pairwise agreement (MPA) of 67.7% in a three-class task of word assessment in customer reviews (NB: α and MPA are not directly comparable). In text annotation task we obtained α=0.278 for all texts and 0.312 for the texts that got non-zero scores (five-class). Hsueh et al (2009) report MPA among Amazon Mechanical Turk annotators to be 35.3% for a four-class task of political blog posts annotation. Ku et al (2006) claim to reach a much higher agreement of 64.7% for four-class blog annotation and 41.2% for news annotation by specially selected assessors. Nevertheless, none of these levels is impressively high. In relation to this, Hsueh and colleagues (2009) point at the problem of political blogs' ambiguity. We tend to agree that this ambiguity and general lack of societal consensus on the polarity of political issues, not (or at least not only) the lack of quality, cause the low agreement. Therefore, disagreeing individuals cannot be filtered out because they may reflect an important part of the public opinion spectrum. A milder measure of divergence of an annotator's mean score from the global mean allows for a lot of disagreement on individual items. It shows that in our case only 0.5% of all annotations were made by individuals strongly deviating from the global mean.

4. Results

4.1. Word and text assessment results

The majority of words (4,753) were annotated as neutral and therefore excluded from the lexicon. Table 1 shows that although negatively assessed words prevail, positive words have been also detected. At the same time, highly emotional words are quite a few.

Table 1. Distribution of mean scores over words

Mean score (rounded)	Number of words with such score	Share of words with such score, %
−2	225	3
−1	1,666	22
0	4,753	63
1	853	11
2	49	0.6

We have also calculated the variance of scores for each word. Although, as mentioned above, disagreement not necessarily indicates low quality, the usefulness of highly disputable words for sentiment classification of texts is doubtful. Since distance between two neighboring values of scores is one, we have regarded all words with variance =>1 as candidates for being discarded. However, we have found only 153 such words, and most of them looked like sentiment-bearing. In some cases their polarity seemed quite obvious: e.g. gorgeous (*сногсшибательный*), filth (*мерзость*), long-suffering (*многострадальный*), first-class (*первосортный*), while others looked ambiguous: e.g. endless (*бесконечный*), quality (*качество*), and tolerance (*терпимость*). At this stage of the research they have been included in the lexicon with their mean scores (since their number is anyway negligibly small).

The distribution of scores over texts is similar to that of words (see Table 2). Most texts were marked as neutral. Positive class size is obviously insufficient (it relates to the negative class as 1:4.6). The same unbalanced class structure in political blogs is also pointed at by Hsueh et al (2009).

Table 2. Distribution of mean scores over texts

Mean score (rounded)	Number of texts with such score	Share of texts with such score
−2	75	0.4
−1	6,546	34
0	11,760	61
1	1,427	7
2	23	23

4.2. Lexicon quality evaluation

After neutral word filtering and leaving out the words that did not occur in the relevant texts, our lexicon comprised 2,753 items. We installed this lexicon into *SentiStrength* freeware for quality evaluation (Thelwall et al, 2010). All texts were lemmatized with *MyStem2* (Segalovich, 2003) prior to SA. In the default mode, *SentiStrength* ascribes two sentiment scores to each text: one corresponds to the maximal negative word score in the text, and the other—to the maximal positive score. If booster words occur before a given word, the absolute value of its score is increased. If negation is used, the sign of the score is reversed. The integrated text score was then calculated as the mean of the two *SentiStrength* scores.

We defined lexicon quality as the share of correctly detected cases. We first calculated the absolute difference between the rounded mean assessors' score of a given text and the rounded integrated score based on *SentiStrength* results. Then we obtained the share of the exact matches, as well as the share of ±1 class matches, as offered by *SentiStrength* developer (Thelwall et al, 2010). A ±1 class match means that if a text is ascribed to one of the two neighboring classes, the

class is considered correctly predicted. In our case the share of ±1 class matches comprises 93.0% which is comparable to Thelwall's results—96.9% (Thelwall et al, 2010). Prediction of the negative classes is better than that of the positive ones (95% and 59% for '−1' and '−2' classes vs. 82% and 19% for '+1' and '+2' classes). As it can also be seen, moderate classes are predicted much better than extreme classes, which are very small, while the dominant '0' class yields 99.6% of ±1 class matches.

SA systems for Russian use different evaluation techniques. The closest to our case was the *ROMIP* SA competition held on texts from political news and from blogs containing customer opinions (Chetviorkin, Loukachevitch, 2013). As sentiment lexicons are domain sensitive, it would be unfair to directly test our lexicon on the texts of a different type and to compare it to the approaches that were developed specially for this type. It would be equally unfair to apply the *ROMIP* methods to our collection. We therefore performed an indirect comparison of the results, using the same methodology of quality evaluation as *ROMIP*. Its best participants in a three-class blog classification task exceeded their baseline by 12–27% in terms of recall and by 5–29% in terms of precision. In news classification task the respective values were 23–28% and 43–49%. Having converted our data into three classes (positive, negative and neutral), we calculated our baseline, precision and recall (see Table 3).

Table 3. Three-class classification quality

	Recall (macro)	**Precision (macro)**
Our lexicon	0.43	0.44
Baseline	0.33	0.18
Difference	0.10	0.26

The quality of our lexicon is comparable to that of the *ROMIP* approaches used in the blog classification task and is lower than the quality reached for news. It should be noted that class distribution of the *ROMIP* news collection was much more balanced (Panicheva, 2013) than that of both its blog collection and of our sample. This has made the task of exceeding the baseline more difficult in blog SA. In contrast to most *ROMIP* methods, our lexicon is publicly available and may be improved by the research community.

5. Conclusion and future research

We have presented a lexicon for sentiment analysis of political and social Russian-language blogs. Its quality is comparable to the results obtained for English-language Twitter and for Russian-language blogs with customer opinions. We have also described the results of words and texts annotation based on a crowdsourcing approach. The lexicon and the annotated collection are publicly available at our website linis-crowd.org that allows further crowdsourcing of sentiment markup. This web

resource is aimed at the widest research community. While the lexicon can be already used by social scientists, the collection may serve as a benchmark for testing new sentiment instruments. In particular, we are now using it for training machine learning SA algorithms that should help increase the quality of SA. We also plan to improve the lexicon by replicating our research on a collection of blog comments that are potentially much more emotional.

6. Acknowledgements

This work was supported by the Russian Foundation for Humanities, project 'Development of a publicly available database and a crowdsourcing website for testing sentiment analysis instruments', Grant No 14-04-1203.

References

1. *Alexeeva S., Koltsova E., Koltcov S.* (2015) Linis-crowd.org: A lexical resource for Russian sentiment analysis of social media [Linis-crowd.org: lexicheskij resurs dl'a analiza tonal'nosti sotsial'no-politicheskix tekstov], Computational Linguistics and computantional ontologies: Proceedings of the XVIII joint Conference "Internet and modern society (IMS-2015)" [Kompyuternaya lingvistika i vyichislitelnyie ontologii: sbornik nauchnyih statey. Trudyi XVIII ob'edinennoy konferentsii «Internet i sovremennoe obschestvo» (IMS-2015)], St. Peterburg, pp. 25–34.
2. *Bodrunova S., Koltsov S., Koltsova O., Nikolenko S., Shimorina A.* (2013) Interval Semi-Supervised LDA Classifying Needles in a Haystack, Proceeding of the 12th Mexican International Conference on Artificial Intelligence (MICAI 2013) Part I: Advances in Artificial Intelligence and Its Applications, Berlin: Springer Verlag, pp. 265–24.
3. *Chetviorkin I. I., Braslavski P. I., Loukachevitch N. V.* (2012), Sentiment Analysis Track at ROMIP 2011, Proceedings of International Conference Dialog, pp. 739–746.
4. *Chetviorkin I., Loukachevitch N.* (2012) Extraction of Russian Sentiment Lexicon for Product Meta-Domain, Proceedings of COLING 2012: Technical Papers, pp. 593–610.
5. *Chetviorkin I., Loukachevitch N.* (2013) Sentiment Analysis Track at ROMIP 2012, Proceedings of International Conference Dialog, Vol. 2, pp. 40–50.
6. *Esuli A., Sebastiani F.* (2006) SentiWordNet: A publicly available lexical resource for opinion mining, Proceedings of 5th International Conference on Language Resources and Evaluation (LREC), Genoa, pp. 417–422.
7. *Godbole N., Srinivasaiah M., Skiena S.* (2007) Large Scale Sentiment Analysis for News and Blogs, ICWSM'2007, Boulder, Colorado, USA.
8. *Hong Y., Kwak H., Baek Y., Moon S.* (2013) Tower of Babel: a crowdsourcing game building sentiment lexicons for resource-scarce languages, Proceedings of the 22nd International World Wide Web Conference (WWW), pp. 549–556.

9. *Hsueh P., Melville P., Sindhwani V.* (2009) Data quality from crowdsourcing: a study of annotation selection criteria, Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing, Boulder, Colorado, pp. 27–35.
10. *Hu M., Liu B.* (2004) Mining and summarizing customer reviews, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Seattle, WA, pp. 168–177.
11. *Koltsova O., Koltcov S., Alexeeva S.* (2014) Do ordinary bloggers really differ from blog celebrities? Proceedings of WebSci ‹14 ACM Web Science Conference, Bloomington, IN, USA, NY: ACM, pp. 166–170.
12. *Koltsova O., Shcherbak A.* (2015) 'LiveJournal Libra!': The political blogosphere and voting preferences in Russia in 2011–2012, New Media and Society, vol. 17, no. 10, pp. 1715–1732.
13. *Ku L.-W., Liang Y.-T., Chen H.-H.* (2006) Opinion Extraction, Summarization and Tracking in News and Blog Corpora, Proceedings of the AAAI-CAAW›06.
14. *Loukachevitch N. V., Blinov P. D., Kotelnikov E. V., Rubtsova Y. V., Ivanov V. V., Tutubalina E.* (2015) SentiRuEval: testing object-oriented sentiment analysis systems in Russian, Proceedings of International Conference Dialog, Vol. 2.
15. *Medhat W., Hassan A., Korashy H.* (2014) Sentiment analysis algorithms and applications: a survey, Ain Shams Engineering Journal, Vol. 5, Issue 4, pp. 1093–1113.
16. *Mihalcea R., Banea C., Wiebe J.* (2007) Learning multilingual subjective language via cross-lingual projections, Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, pp. 976–983.
17. *Mohammad S, Dorr B., Hirst G., Turney P.* (2011) Measuring degrees of semantic opposition, Technical report, National Research Council Canada.
18. *Mohammad S. M., Turney, P. D.* (2013), Crowdsourcing a word-emotion association lexicon, Computational Intelligence, Vol. 29 no. 3, pp. 436–465.
19. *Morkovkin V. V.* (2003) Explanatory dictionary of Russian language: structural words: prepositions, conjunctions, particles, interjections, parentheses, pronouns, numbers, connections [Ob'jasnitelnyj slovar' russkogo jazyka: Structurnyje slova: predlogi, sojuzy, chastitsy, mezhdometija, vvodnyje slova, mestoimenija, chislitelnyje, svjazannyje slova], Astrel, Moscow.
20. *Nikolenko S., Koltcov S., Koltsova O.* (2015) Topic Modeling for Qualitative Studies. Journal of Information Science (R&R).
21. *Pang B., Lee L.* (2004) A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, 42nd Meeting of the Association for Computational Linguistics[C] (ACL-04), pp. 271–278.
22. *Pang B., Lee L.* (2008) Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval, Vol. 2, no. 1–2, pp. 1–135.
23. *Panicheva P.* (2013) ATEX: a rule-based sentiment analysis system processing texts in various topics [Sistema sentimentnogo analiza ATEX osnovannaya na pravilah pri obrabotke tekstov razlichnyh tematik], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2013" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2013"], pp. 101–113.

24. *Segalovich I.* (2003) A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine, Proceedings of MLMTA'2003, pp. 273–280.
25. *Steyvers M., Griffiths T.* (2004) Finding scientific topics, Proceedings of the National Academy of Sciences, Vol. 101, no. Suppl. 1, pp. 5228–5235.
26. *Taboada M., Brooke J., Tofiloski M., Voll K., Stede M.* (2011) Lexicon-Based Methods for Sentiment Analysis, Computational Linguistics, Vol. 37, no. 2, pp. 267–307.
27. *Thelen M., Riloff E.* (2002) A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts, Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), pp. 214–221.
28. *Thelwall M., Buckley K., Paltoglou G., Cai D., Kappas A.* (2010) A. Sentiment strength detection in short informal text, Journal of the American Society for Information Science and Technology, Vol. 61, no. 12, pp. 2544–2558.
29. *Turney P.* (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, Proceedings of 40th Meeting of the Association for Computational Linguistics, Philadelphia, PA, pp. 417–424.

IMPROVING DISTRIBUTIONAL SEMANTIC MODELS USING ANAPHORA RESOLUTION DURING LINGUISTIC PREPROCESSING

Koslowa O. (evezhier@gmail.com)
National Research University Higher School of Economics, Moscow, Russia

Kutuzov A. (andreku@ifi.uio.no)
University of Oslo, Norway

In natural language processing, distributional semantic models are known as an efficient data driven approach to word and text representation, which allows computing meaning directly from large text corpora into word embeddings in a vector space. This paper addresses the role of linguistic preprocessing in enhancing performance of distributional models, and particularly studies pronominal anaphora resolution as a way to exploit more co-occurrence data without directly increasing the size of the training corpus.

We replace three different types of anaphoric pronouns with their antecedents in the training corpus and evaluate the extent to which this affects the performance of the resulting models in lexical similarity tasks. CBOW and SkipGram distributed models trained on Russian National Corpus are in the focus of our research, although the results are potentially applicable to other distributional semantic frameworks and languages as well. The trained models are evaluated against RUSSE'15 and SimLex-999 gold standard data sets. As a result, we find that models trained on corpora with pronominal anaphora resolved perform significantly better than their counterparts trained on baseline corpora.

Keywords: anaphora resolution, distributional semantics, word2vec, semantic similarity, vector space models, neural embeddings

РАЗРЕШЕНИЕ АНАФОРЫ В ОБУЧАЮЩЕМ КОРПУСЕ КАК СПОСОБ УЛУЧШЕНИЯ КАЧЕСТВА ДИСТРИБУТИВНО-СЕМАНТИЧЕСКИХ МОДЕЛЕЙ

Козлова О. (evezhier@gmail.com)
Национальный Исследовательский Университет Высшая Школа Экономики, Москва, Россия

Кутузов А. (andreku@ifi.uio.no)
Университет Осло, Норвегия

Ключевые слова: разрешение анафоры, дистрибутивная семантика, word2vec, семантическая близость, векторные репрезентации лексики, искусственные нейронные сети

1. Introduction

Natural language semantics is the level of human language which is least formalized and understood: there is no general agreement even on what exactly to consider the meaning of a language unit. Thus, modeling it computationally is a very ambitious task. Distributional semantics is based on the hypothesis expressed in [Firth 1957] that meaning is composed of typical contexts in which a given unit occurs. This conceptualization of meaning can be represented with a vector space induced from large and representative corpora (with contexts as dimensions and frequencies as numeric values of components), so semantics becomes computationally tractable.

This approach to meaning representation has been studied for several decades. However, important events happened in 2013, when prediction-based models became popular, particularly Continuous Bag-of-Words (CBOW) and Continuous SkipGram algorithms proposed in [Mikolov et al. 2013] and implemented in *word2vec* software tool. Dense vectors generated by such models are called 'embeddings', and they are induced using machine learning techniques, such as artificial neural networks.

These algorithms were able to dramatically reduce computational complexity of training vector semantic models, without compromising performance [Baroni et al. 2014]. Additionally, they showed some interesting properties, such as geometrical vector operations reflecting semantic relations. This quickly resulted in these algorithms becoming an established standard in the field, used in wide spectrum of natural language processing applications. They are employed in constructing semantic maps of language and high-level processing, such as representing phrase and text semantics. This in turn forms a basis for various practical tasks: sentiment analysis, machine translation, information retrieval, fact extraction and natural language generation.

In terms of neural word embeddings, researchers traditionally search for ways of performance improvement in polishing learning algorithms and optimizing hyperparameters. However, in the presented research we deal with linguistic preprocessing of training corpora. Particularly, we describe an experiment of merging achievements of computational discourse analysis and distributional semantics. We address linguistic phenomenon of anaphora and treat its automatic resolution as a means to disclose 'under-the surface' word contexts to supply more training material for distributional models.

The experiment described further deals with three types of pronominal anaphors: relatives, reflexives and personal pronouns. As a rule, when training a distributional model, these entities are considered to be stop words and discarded during corpora preprocessing. We argue, instead, that this leads to waste of training resources, and models would benefit from associating proforms (elements with dependent semantic interpretation) with their antecedents. We show that models trained on Russian National Corpus [Plungian 2005] (further RNC) with resolved pronominal anaphora perform significantly better in standard lexical similarity tasks.

The structure of the paper is as follows: first, we present a brief overview of known approaches to linguistic preprocessing in distributional semantics, clarify related terminology and account for the state-of-the-art anaphora resolution for Russian. Next, we describe setting of our experiment, tools and evaluation methods. Then we discuss experimental results and outline the directions for the future work.

2. Related work

The topic of linguistic preprocessing of training corpora for distributional models is not novel. In [Pekar, 2004] some possible tips (lemmatization, morphological and syntactic analysis, rare context words removal) are explored for English. Lemmatization is considered most valuable, as apart from increasing model performance, it also reduces feature space. As stated in [Baroni, 2008], the preprocessing pipeline largely depends on availability of resources for a particular language and their quality, which may be an obstacle for advanced linguistic analysis. Tokenization, lemmatization and POS-tagging are considered to be baseline stages.

As for the connection between distributional semantics and discourse analysis, applying vector models to anaphora and coreference resolution is a research topic which draws an increasing amount of attention. Competitive results have already been demonstrated, as in [Clark, 2015]. However, to the best of our knowledge, the extent to which anaphoric associations could enrich the quality of the models themselves has been largely unexplored both for English and Russian.

The most relevant experiment to the one described in the present paper involved coreference chains mined from a large corpus as a training resource for a model used for semantic tasks [Adel, Schütze 2014]. This resulted in better precision in detecting antonyms. The authors acknowledge coreference chains to be valuable as a supplementary resource for creating semantic representations. However, it is reported that coreference chains alone encompass "only a small subset of word-word relations encoded in raw text". Our workflow addresses this issue.

3. Linguistic preprocessing of training corpora

Although it is possible to construct representations directly from tokenized text corpora, natural language texts are full of complexities, which may eventually cause significant adverse effect on the results. Considering this, performing minimal linguistic preprocessing is standard practice before training a distributional model. Constituent stages depend on many factors including the nature of the primary task, and may be very specific. However, there exists a number of common procedures implemented before the model actually starts learning vectors for words.

3.1. Tokenization

To model meaning through distribution, one first determines semantics of which language units is of interest. Thus, token segmentation is the only preparatory step required by vector models themselves. Minimal additional preprocessing starts when tokens are converted to lowercase, which is not obligatory but improves performance by eliminating the difference between capitalized and non-capitalized word forms. Note that some well-known publicly available models for English (for example, Google News model released along with *word2vec* code) lack even this basic preprocessing.

3.2. Morphological analysis: lemmatization and PoS-tagging

Replacing word tokens with their lemmas (normal forms or word types) has a two-fold purpose: to reduce vocabulary size and to 'squeeze' all word forms co-occurrence information into one vocabulary item. This is especially important for rich morphology languages like Russian, where each noun may have at least 12 word forms, and having a separate vector for each of them is usually impractical (except when one needs to exploit relations between different grammatical forms of the same word). This stage involves specific linguistic tools, since sloppy analysis may even decrease vector quality. Sometimes lemmas are additionally supplied with part-of-speech tags, which helps to resolve ambiguity where lemmatization alone does not suffice [Kutuzov, Andreev 2015]. For example, Russian word "*печь*" can be interpreted either as a noun (*stove*) or as a verb (*to bake*). Without PoS information, a model would try to learn one vector for both words, which obviously does not make sense. At the same time, adding PoS tags makes it two words: "*печь_S*" and "*печь_V*", which then acquire two different vector representations.

3.3. Stop words

Some lexical units are considered to be not very useful in semantic tasks, and thus are filtered out from the corpus before training to get rid of unwanted 'noise'[1]. Such stop words are divided into two large groups:

1. **Functional words** (prepositions, numbers, conjunctions, etc). They do not possess their own meaning, and thus it does not make sense to spend training time on them.
2. In some tasks, it is useful to filter out **very frequent** (maybe, domain-specific) words or to downsample them during training so that their influence were limited.

Pronouns are as a rule considered to be examples of the first stop words type. Note that in the presented research we do not remove them from the training corpus, but replace them with their antecedents. In the next section we describe the role of pronominal anaphora in training corpora and give an account of anaphora resolution systems for Russian.

4. Pronominal anaphora and its resolution for Russian

Anaphora is a linguistic phenomenon whereby the interpretation of an occurrence of one expression depends on the interpretation of an occurrence of another or whereby an occurrence of an expression has its referent supplied by an occurrence of some other expression in the same or another sentence [King, Jeffrey C. 2013]. It is closely associated and sometimes confused with coreference. In both cases two

[1] Of course, it depends on the nature of the task. For example, in authorship attribution removing stop words can be undesirable.

or more expressions have the same referent, but the latter is much broader and does not presuppose interpretation dependency. For instance:

(1) a) *Мальчик$_i$, который$_i$ сидел за столом, был задумчив.*
The boy$_i$ who$_i$ was sitting at the table was lost in thought.
b) *Вася$_i$ сидел за столом. Мальчик$_i$ был задумчив.*
Vasya$_i$ was sitting at the table. The boy$_i$ was lost in thought

a) is the case of anaphora, since it is not possible to assign semantic interpretation to *'который'* ('who') without considering its co-indexical *'Мальчик'* ('The boy'). The unit which is dependent in its interpretation is called the **anaphor**, while the one which provides interpretation is the **antecedent**.
b) is the case of so-called coreferential noun phrases, each co-indexed unit possesses its own lexical meaning, but they all refer to the same person.

In this research, we limit ourselves to anaphora, leaving coreference for future work.

Anaphoric pronouns occupy top positions in frequency lists for most languages. Russian is no exception. Table 1 provides frequency statistics for 3 pronominal anaphora types we enumerated earlier, in 2 well-known Russian corpora after PoS-tagging.

Table 1. Frequencies of anaphoric pronouns (instances per million)

Corpus	Relatives	Reflexives	Personal pronouns
RNC[2]	4,264.8	6,486.2	31,377.7
Open Corpora[3]	5,891.4	5,367.3	18,961.0

It is obvious that these pronouns are in fact coreferential representations of meaningful words. They convey additional information on the distribution of the latter in a covert, non-explicit way.

Our major hypothesis is that considering the huge amount of pronominal anaphors in natural language texts, the standard practice of discarding them as stop words results in loss of a significant number of distributional contexts, while resolving anaphora would provide a model with more contexts for words functioning as antecedents, optimizing data usage. This is particularly important for relatively small corpora, such as Russian National Corpus. Note that it is regularly used as a primary resource for developing computational tools for the Russian language, including distributional models.

In natural language processing, to perform anaphora resolution means to associate co-indexed elements with each other, finding their common referent. In other words, this is a task of finding the most probable antecedent for a given anaphor. It is a well-studied field of natural language processing with its own methods and

[2] http://ruscorpora.ru/en

[3] http://opencorpora.org

tools. Unfortunately, there is evident lack of corresponding publicly available tools for Russian. For this experiment we used *An@phora*[4], an open-source tool for pronominal anaphora resolution, which has been successfully tested on the major pronominal anaphor types: relatives, reflexives, possessives and personal pronouns [Kutuzov, Ionov 2014]. It is based on a set of rules and takes as an input morphologically analyzed sentences, looking for candidate antecedents in a window of *n* words length to the left of the current anaphor. The authors claim that *An@phora* achieves precision and recall of up to 0.6 on Russian texts.

5. Experimental setting

5.1. Resources and tools

To test our core assumption, we use full Russian National Corpus as training material. It was linguistically pre-processed with *NLTK* [Bird et al. 2009], *Mystem* [Segalovich 2003] and *An@phora* resolver. To train neural embedding models on the resulting corpora, we employed *Gensim* framework [Řehůřek and Sojka 2010], which implements CBOW and SkipGram algorithms in Python.

5.2. Anaphora resolution

Our aim is to explore the overall effect of anaphora resolution on the resulting vector models, along with type-wise comparison. To this end, antecedents for all anaphors in the training corpus were found. Then, all types of anaphors were replaced with the detected antecedents. Additionally, we produced 3 more transformed variants of each document, in which only one of 3 anaphora types recognized by our resolver was replaced. The exact list of anaphors identified by *An@phora* is given in Table 2.

Table 2. Types of Russian anaphoric pronouns detected by *An@phora*

Personal pronouns	Relatives	Reflexives
1. *он* 2. *она* 3. *оно* 4. *они* 5. *его* 6. *ее* 7. *их* 8. *мой*	1. *который*	1. *себя* 2. *свой*

[4] http://ling.go.mail.ru/anaphora

The length of analysis window (the distance at which an antecedent was searched) was set to 23 words, as per recommendations of *An@phora* authors. Table 3 provides statistics of anaphora replacement performed in the Russian National Corpus.

Table 3. Total number of anaphors and resolved anaphors in RNC

Anaphor type	Total occurrences	Number of resolved anaphors (antecedent found)
Reflexives	1,200,079	1,105,853 (92.15%)
Relatives	789,080	743,018 (94.16%)
Personal pronouns	5,805,519	4,412,671 (76.00%)

As a result, the majority of anaphoric pronouns were replaced with their antecedents (mostly nouns and noun phrases). See the example below, with the original sentence as *(a)*, and the resulting one as *(b)*:

(2) a) *Речь идёт о том, что* ***коллектив***$_k$ *должен нести ответственность за результаты* ***своей***$_k$ *деятельности и выступать продавцом* ***своих***$_k$ *услуг на рынке.*
The matter is that a company$_k$ *must take responsibility for the results of* ***its***$_k$ *activity and act as a vendor of* ***its***$_k$ *services.*
b) *Речь идёт о том, что* ***коллектив*** *должен нести ответственность за результаты* ***коллектив*** *деятельности и выступать продавцом* ***коллектив*** *услуг на рынке.*
The matter is that a ***company***$_k$ *must take responsibility for the results of* ***company***$_k$ *activity and act as a vendor of* ***company***$_k$ *services.*

The example illustrates how anaphora was successfully resolved for the word *'свой'* ('its'). Note that the difference in word forms is eliminated during lemmatization. As a result, the model learned closer connection between the words *'коллектив'* ('company'), *'услуги'* ('services') and *'деятельность'* ('activity').

In case no antecedent was found for the given anaphor, the latter was discarded as a stop word. Less than 10% of all texts in the corpus (mainly, short news bulletins and jokes) contained no anaphoric relations. Note that the sentences with resolved anaphora replaced the original ones, so the raw amount of training material did not change significantly.

5.3. Preprocessing

Four training corpora were produced during the previous stage: one for each of the 3 anaphora types and one with all identified anaphors replaced. The 5th one is the baseline control corpus with no anaphors replaced. Preprocessing included sentence splitting (with *NLTK*), tokenization, lemmatization, PoS-tagging (with *Mystem*) and stop words removal. For the latter we used NLTK default stop list for Russian. Numeric tokens and punctuation were also discarded.

5.4. Models training

We trained distributional models with the following fixed hyperparameters:

- Vector size: 300;
- Minimal frequency to consider a word during training: 3;
- Negative samples: 15.
  They were chosen as the best for RNC models ([Kutuzov, Andreev 2015]).

We experimented with tuning the following hyperparameters:

1. Learning algorithms: SkipGram or CBOW;
2. Width of symmetrical context window: 1, 2, 3, 5, 10, or 20 words.

As we have 5 training corpora (control `baseline' corpus, all anaphors replaced, only personal pronouns replaced, only relatives replaced and only reflexives replaced), 2*6*5=60 models were trained all in all.

6. Evaluation

6.1. Methods

The task of measuring semantic relatedness is fundamental for distributional semantics and is a good test for the results of our experiment. If word vectors trained on corpora with anaphora resolved gained consistent increase in quality, it means that the experiment was successful and anaphora resolution does increase the models' performance. The results were evaluated against two gold standards for measuring semantic similarity in word pairs: **RUSSE'15** training set [Panchenko et al. 2015] and **Simlex-999** [Hill et al. 2015].

In the first case using the training dataset instead of the test dataset allows for more sound statistics due to larger number of word pairs with annotated measure of relatedness: 209,320 versus 14,836 in the test set. RUSSE evaluation standard suggests four tasks: **hj** (Spearman's correlation with expert annotations), **rt** (average precision for word pairs from *RuThes Lite*), **ae** (average precision for associations from the *Russian Associative Thesaurus*) and **ae2** (average precision for associations from the *Sociation.org*).

6.2. Results

In Table 4 we show the difference between the best baseline models trained on raw RNC (no anaphora resolution) and those RNC versions where anaphora was resolved. We also provide training parameters for every model (learning algorithm and window size). Full tables with all the results can be found in the Appendix[5].

[5] http://ling.go.mail.ru/misc/dialogue_2016.html

Table 4. Performance in RUSSE tasks for the best experimental models

RUSSE task	Best raw models with lemmatization and PoS-tagging	Best anaphora-enriched models
hj	0.75608 CBOW / Window 10	**0.76529** Reflexives replaced CBOW / Window 20
rt	0.78311 CBOW / Window 1	**0.78348** Relatives replaced CBOW / Window 1
ae	0.83045 SkipGram / Window 20	**0.83899** All anaphors replaced SkipGram / Window 20
ae2	0.85341 CBOW / Window 20	**0.86660** All anaphors replaced CBOW / Window 20

Improvements are obvious: in all tasks, the models trained on corpora with anaphora resolution were the best. The most significant benefit was seen in **hj** and associative **ae2** tasks. Relatedness task **rt** demonstrated the least evident change. In most tasks CBOW algorithm worked best while in **ae** Skip-gram algorithm was a better option. However, anaphora resolution everywhere resulted in stable, parameter-independent quality growth. We could not single out one best combination of training parameters and anaphora type to resolve in corpora, so it seems to be task-dependent.

Additionally, we evaluated all the models which performed best in the RUSSE tasks against Simlex-999. This is the Russian section of multilingual human-annotated dataset, consisting of 999 word pairs manually annotated with semantic similarity. The results are summarized in Table 5.

Table 5. Spearman correlation against SimLex-999 gold standard

Model	Spearman correlation
All anaphors replaced CBOW / window 20	**0.613**
Raw corpus CBOW / window 20	0.606
Reflexives replaced CBOW / window 20	0.604
All anaphors replaced SG / window 20	0.599
Raw corpus CBOW / window 10	0.597
Raw corpus SG / window 10	0.589

Even this relatively sparse test data confirms the advantage which anaphora resolution lends to distributional models. It is interesting that resolution of personal pronouns alone did not result in high accuracy. In our opinion, this is a backlash of relatively low quality of our current anaphora resolver. The majority of wrong antecedents were assigned to personal pronouns. Along with their high frequency, this significantly impaired the results.

Resolution of all anaphors gave boost to two associative metrics, which normally benefit from large window parameter. Anaphora resolution generally increases the number of words in sentences, thus it is natural that large windows also allowed anaphora to take greater effect. The same explanation can be offered for relatively small improvement in case of **rt**, where the best models were those with the smallest window possible.

Table 6 and Table 7 demonstrate the performance of 4 best models trained on raw corpora with minimal preprocessing (only lower-casing tokens). This metrics illustrates the extent to which linguistic analysis as a whole improves model performance.

Table 6. RUSSE metrics of the best models trained on raw corpus with minimal preprocessing

RUSSE task	CBOW / window 1	CBOW / window 10	CBOW / window 20	SG / window 20
hj	0.26680	0.47501	0.54553	0.58660
rt	0.63063	0.71158	0.71750	0.67318
ae	0.54013	0.66039	0.67387	0.67318
ae2	0.54013	0.78991	0.80959	0.67318

Table 7. Spearman correlation against SimLex-999 gold standard (the best models trained on raw corpus with minimal preprocessing)

CBOW / window 1	0.2538
CBOW / window 10	0.3982
CBOW / window 20	0.4311
SG / window 20	0.5093

7. Conclusions and future work

In this paper we employed pronominal anaphora resolution as a way to optimize data usage for training word embedding models. In the course of our experiments, we trained neural distributional models on Russian National Corpus with resolved anaphors of the following types: personal pronouns, relatives, and reflexives. The performance of the resulting models in various semantic tasks was then evaluated. We showed that anaphora resolution results in evident improvement of models' performance. With the same input and identical set of hyperparameters, 'anaphora-enriched' models were consistently ahead their baseline competitors by several points.

These results are especially promising, considering that the employed anaphora resolution tool (*An@phora*) is not perfect: with 60% accuracy, a good part of the detected antecedents is wrong, and for some anaphors, antecedents are not found at all. Nevertheless, a consistent increase in semantic similarity task performance is observed.

Unfortunately, no other publicly available anaphora resolution tools for Russian exists yet. Thus, we were unable to estimate how much the resolver performance influences the results. We look forward to new upcoming tools, which would make such an experiment possible.

With corresponding instruments at hand, it would also be interesting to resolve not only anaphora but also coreference chains, to check if it provides additional performance boost. We assume this a more advanced problem, since mere replacement will not work: getting new contexts for some units will mean losing those for others.

Anaphora is a universal linguistic phenomenon, so our work can be applied to other languages as well. More tools for anaphora resolution are available for some languages, cf. for example, Stanford Deterministic Coreference Resolution System [Recasens et al. 2013]. Thus, experiments on English are in our nearest plans.

References

1. *Adel H., Schütze H.* (2014), Using Mined Coreference Chains as a Resource for a Semantic Task. In EMNLP (pp. 1447–1452).
2. *Baroni, M.* (2008). Distributional Semantics (slides)
3. *Baroni M., Dinu G., Kruszewski, G.* (2014), Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 238–247.
4. *Bird S. Loper E., Klein E.* (2009), Natural Language Processing with Python. O'Reilly Media Inc.
5. *Kevin Clark* (2015), Neural coreference resolution. Stanford Report
6. *Firth J. R.* (1957), "A synopsis of linguistic theory 1930–1955". Studies in Linguistic Analysis (Oxford: Philological Society): 1–32. Reprinted in F.R. Palmer, ed. (1968). Selected Papers of J.R. Firth 1952–1959. London: Longman.
7. *Hill F., Reichart R., Korhonen A.* (2015), Simlex-999: Evaluating semantic models with (genuine) similarity estimation. Computational Linguistics.
8. *King Jeffrey C.* (2013), "Anaphora", The Stanford Encyclopedia of Philosophy (Summer 2013 Edition), Edward N. Zalta (ed.).
9. *Kutuzov A., Ionov M.* (2014), The impact of morphology processing quality on automated anaphora resolution for Russian, Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue" (Bekasovo, June 4–8, 2014), issue 13 (20) , Moscow, RGGU.
10. *Kutuzov A., Andreev I.* (2015), Texts in, meaning out: Neural language models in semantic similarity tasks for Russian, Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue" (Moscow, May 27–30, 2015), issue 14 (21), Moscow, RGGU.

11. *Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J.* (2013), Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems.
12. *Panchenko A., Loukachevitch N. V., Ustalov D., Paperno D., Meyer C. M., Konstantinova N.* (2015), RUSSE: The First Workshop on Russian Semantic Similarity, Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue". (Moscow, May 27–30, 2015), issue 14 (21), Moscow, RGGU.
13. *Pekar, V.* (2004, August). Linguistic preprocessing for distributional classification of words. In Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries (pp. 15–21). Association for Computational Linguistics.
14. *Plungian V. A.* (2005), Why we make Russian National Corpus? [Зачем мы делаем Национальный корпус русского языка?], Otechestvennye Zapiski, 2
15. *Řehůřek R., Sojka P.* (2010), Software framework for topic modeling with large corpora, Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta.
16. *Segalovich I.* (2003), A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine, MLMTA, pp. 273–280.
17. *Recasens, M., de Marneffe, M. C., & Potts, C.* (2013). The Life and Death of Discourse Entities: Identifying Singleton Mentions, HLT-NAACL (pp. 627–633).

MANUALLY CREATED SENTIMENT LEXICONS: RESEARCH AND DEVELOPMENT

Kotelnikov E. V. (kotelnikov.ev@gmail.com),
Bushmeleva N. A. (bushmeleva_na@list.ru),
Razova E. V. (razova.ev@gmail.com),
Peskisheva T. A. (peskisheva.t@mail.ru),
Pletneva M. V. (pletneva.mv.kirov@gmail.com)

Vyatka State University, Kirov, Russia

The sentiment lexicons are an important part of many sentiment analysis systems. There are many automatic ways to build such lexicons, but often they are too large and contain errors.

The paper presents the algorithm of sentiment lexicons creation for a given domain based on hybrid—manual and corpus-based—approach. This algorithm is used for the development of the sentiment lexicons by means of four human annotators each for five domains—user reviews of restaurants, cars, movies, books and digital cameras. Created sentiment lexicons are analyzed for inter-annotator agreement, parts of speech distribution and correlation with automatic lexicons.

The performance of the sentiment analysis based on the created sentiment lexicons is researched and compared with the performance of the existing sentiment lexicons. The experiments with text corpora on various domains based on SVM show high quality and compactness of the human-built lexicons.

Key words: sentiment analysis, sentiment lexicons, manual approach, corpus-based approach, SVM

СЛОВАРИ ОЦЕНОЧНОЙ ЛЕКСИКИ, СОЗДАННЫЕ ВРУЧНУЮ: РАЗРАБОТКА И ИССЛЕДОВАНИЕ

Котельников Е. В. (kotelnikov.ev@gmail.com),
Бушмелева Н. А. (bushmeleva_na@list.ru),
Разова Е. В. (razova.ev@gmail.com),
Пескишева Т. А. (peskisheva.t@mail.ru),
Плетнева М. В. (pletneva.mv.kirov@gmail.com)

Вятский государственный университет, Киров, Россия

Ключевые слова: анализ тональности, словари оценочной лексики, экспертный подход, подход на основе корпусов, метод опорных векторов

1. Introduction

In recent years the sentiment analysis is one of the hottest research areas in natural language processing (Liu, 2012). The challenges to the researchers are both theoretical aspects, such as the objective laws of the sentiment expressions in the natural language, and the practical aspects, e. g., the analysis of consumer products and services reviews, the monitoring of social networks, the political studies (Feldman, 2013).

There are two main approaches to the sentiment analysis (Taboada et al., 2011): lexicon-based and machine learning. The first of them determines the text sentiment by means of individual words polarity in the text. The latter considers the task of sentiment analysis as the problem of text categorization. Both approaches require high quality sentiment lexicons: even in the text categorization methods the word weights are often proportional to word polarity and strength.

There are many studies on the problem of sentiment lexicons creating. They generally use three main approaches (Liu, 2012): manual approach, dictionary-based approach, and corpus-based approach.

In the manual approach the sentiment lexicons are constructed by human annotators. In the dictionary-based approach the sentiment lexicons are created with the help of the universal dictionaries and thesauri, e. g., WordNet (Fellbaum, 1998). In the corpus-based approach the sentiment lexicons are built based on the analysis of text corpora. Also the various hybrid combinations of these approaches are used.

Though the problem of sentiment lexicons creation is very important, little attention is paid to the evaluation of the quality and in-depth analysis of the generated lexicons, especially for Russian.

In this paper, firstly, we propose a procedure of creating the sentiment lexicon for a given domain, secondly, we analyze the sentiment lexicon that is constructed by several annotators for various domains, thirdly, we research the performance of these sentiment lexicons in comparison with existing lexicons.

The rest of the paper considers the related work (Section 2) and the used text corpora (Section 3) are considered. At first the corpus-based approach to sentiment words extraction is applied to generate the sentiment lexicons, then their manual annotation is carried out by several annotators (Section 4). The generated lexicons are jointly analyzed (Section 5). Performance of the sentiment analysis based on the sentiment lexicons and Support Vector Machine (SVM) is evaluated (Section 6).

2. Related work

2.1. The creation of sentiment lexicons

Two stages of lexicons creation can be distinguished: 1) the generation of the sentiment-bearing words list, containing the candidates to sentiment lexicon, and 2) the assignment of sentiment labels to these words, e. g. positive/negative/neutral. Both stages are performed either manually or automatically.

Most of the studies on concerning sentiment lexicons creation are carried out on the material of English. For example, Taboada et al. (2011) both stages fulfilled manually. Mohammad and Turney (2013) used the crowdsourcing for the creation of word-emotion and word-polarity association lexicon.

There are also studies for other languages. For example, Amiri et al. (2015) formed word list manually, then this list was annotated by several human annotators by means of web interface.

There are few such studies for Russian. Chetviorkin and Loukachevitch (2012) extracted and weighted sentiment words automatically on the base of machine learning. Manual annotation was performed only for evaluation. Ulanov and Sapozhnikov (2013) built up the lexicons by means of automatic translation of English dictionaries. Blinov and Kotelnikov (2014) created the sentiment lexicon based on the distributed representations of words and used it for aspect-based sentiment analysis. Ivanov et al. (2015) applied the corpus-based approach in the user review domain as well as for aspect-based sentiment analysis.

At present the following sentiment lexicons are publicly available:

- Russian Sentiment Lexicon for Product Meta-Domain (ProductSentiRus)—5,000 words (Chetviorkin, Loukachevitch, 2012)[1];
- NRC Emotion Lexicon translated in Russian via Google Translate (NRC)—4,590 words (Mohammad, Turney, 2013)[2];
- Russian sentiment lexicon—2,914 words (Chen, Skiena, 2014)[3];
- Sentiment lexicon for restaurants domain—7,312 words (including bigrams and trigrams) (Blinov, Kotelnikov, 2014)[4].

These lexicons (except the latter, containing the large part of collocations) are used in our study to compare with the manual lexicons (see Section 6).

2.2. Analysis of lexicons

One of the main purposes of our study is a joint analysis of the word list sentiment labeling. The word list was made by several human annotators for various domains. To our knowledge such in-depth analysis of Russian sentiment lexicons hasn't been performed yet.

Andreevskaia and Bergler (2006) conducted simultaneous labeling of two sentiment lexicons by two teams, which resulted in the high degree of disagreement.

Taboada et al. (2011) compared manual lexicons with dictionaries built using Amazon Mechanical Turk. In addition, a comparison with SentiWordNet was drawn, but only at the level of performance test.

---

1 http://www.cir.ru/SentiLexicon/ProductSentiRus.txt

2 http://www.saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

3 https://sites.google.com/site/datascienceslab/projects/multilingualsentiment

4 http://goo.gl/NhEvWu

As well several sentiment lexicons are compared by the quality of sentiment analysis in English (Musto et al., 2014; Ozdemir, Bergler, 2015) and in Portuguese (Freitas, Vieira, 2013).

Within the context of our study we should mention the work (Kiselev et al., 2015) in which the thorough analysis of 12 existing lexical-semantic resources (printed explanatory dictionaries, dictionaries of synonyms, electronic thesauri) is performed.

3. Text corpora

In our work the reviews of restaurants, cars, movies, books and digital cameras are researched. The reviews of restaurants were collected from the site Restoclub[5], the reviews of cars—from the site Cars@mail.ru[6]. For the rest domains the text corpora of seminar ROMIP2011 and 2012 are used (Chetviorkin et al., 2012; Chetviorkin, Loukachevitch, 2013).

The initial score scales (movies, books, restaurants—ten-point, cameras, cars—fivepoint) were converted to binary scale by the following schemes: for ten-point scale—{1...4} → *neg*, {6...10} → *pos*; for five-point scale—{1...2} → *neg*, {4...5} → *pos*.

As a training set the random chosen ten thousand reviews are used for each domain. For the ROMIP's domains these reviews are chosen from train corpora of ROMIP2011, for remaining domains—from entire corpora. Test sets for ROMIP's domains are equal to the test corpora union of ROMIP2011 and 2012 for each domain separately. As test sets for restaurants and cars all reviews are used except for training reviews.

The characteristics of training and test corpora are given in Table 1.

Table 1. Text corpora (N_{av}—an average number of words per review)

Domain	Train corpora				Test corpora				Total reviews
	Pos	Neg	Total	N_{av}	Pos	Neg	Total	N_{av}	
Restaurants	7,982	2,018	10,000	87	15,353	1,544	16,897	162	26,897
Cars	7,900	2,100	10,000	104	38,148	1,286	39,434	71	49,434
Movies	7,330	2,670	10,000	80	594	126	720	212	10,720
Books	7,888	2,112	10,000	31	356	39	395	235	10,395
Cameras	8,921	1,079	10,000	94	612	54	666	226	10,666

It should be noted that the corpora are highly imbalanced: the part of positive reviews is ranging from 73.3% for the movie training corpus to 96.7% for the car test corpus.

[5] http://www.restoclub.ru

[6] https://cars.mail.ru/reviews

4. Sentiment lexicons creating

The proposed procedure of sentiment lexicon creation consists of three main stages: 1) word weighting and selection; 2) collaborative manual word annotation; 3) consolidation of sentiment lexicons.

At the first stage the morphological analysis of training corpus is performed (we used mystem[7]), then full dictionary of training corpus is formed and stop words are removed. All the words are weighted using the supervised term weighting scheme, e.g., RF (Relevance Frequency), which demonstrated good performance in the text categorization task (Lan et al., 2009). In this scheme the weight of a given word to the sentiment category S is calculated by formula:

$$RF_S = \log_2\left(2 + \frac{a}{\max(1, b)}\right),$$

where a—a number of documents related to category S and containing this word, b—a number of documents not related to category S and containing this word as well.

For each word two weights are calculated: the first weight RF_{pos} towards $S = positive$ and second weight RF_{neg} towards $S = negative$. Two identical word lists, which contain all words from full dictionary, are generated. Lists are sorted, the first—in the order of weights RF_{pos}, and the second—in the order of weights RF_{neg}. First P words from each list are chosen so that $2P = N$, where N—a number of words for manual annotation (at the top of both lists the same words may occur). Thus the dictionary for manual labeling containing N hypothetical sentiment words is made for the second stage.

Table 2 shows the characteristics of full dictionaries and dictionaries for labeling.

Table 2. Size of dictionaries

Domain	Size of full dictionary	Size of labelled dictionary
Restaurants	21,454	10,000
Cars	17,810	10,000
Movies	28,955	10,000
Books	15,328	10,000
Cameras	13,974	10,000

At the second stage M annotators independently label dictionary. In our study $M = 4$ annotators take part in the annotation process. $N = 10{,}000$ is the compromise between the laboriousness and the completeness. The annotators labelled 50,000 words (5 domains) altogether. The dictionary was shuffled before annotation.

Each word can be assigned one of four labels: positive, negative, neutral and unclear. Further neutral words are not used. The unclear word lists are of interest to further studies.

The desktop application that shows the current word, its context and possible labels is used for the labeling process (Fig. 1).

[7] https://tech.yandex.ru/mystem

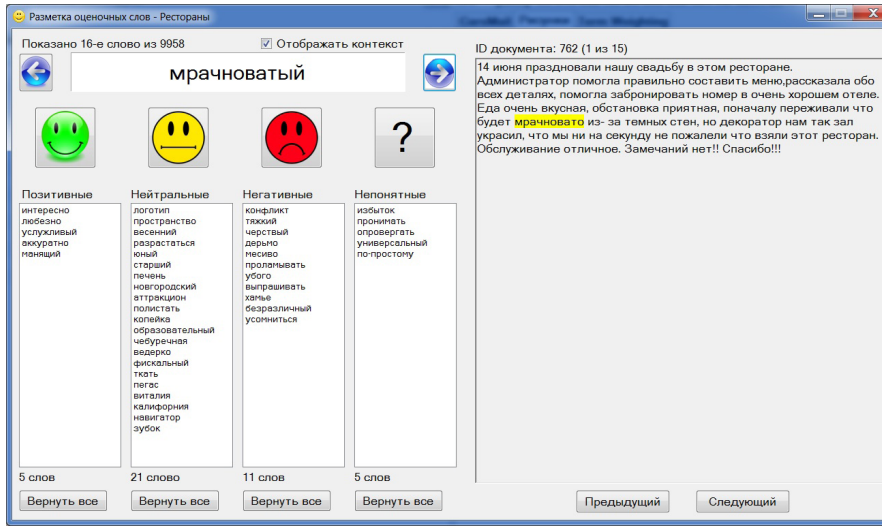


Fig. 1. Annotation tool

The annotators labelled the word as positive or negative in case they could imagine it in any sentiment context of current domain. If the annotator had some doubt the word was labelled as unclear, otherwise as neutral. The average time of labeling of a thousand of words was 90 minutes, overall labeling time was about 300 man-hours.

The anotators had the following main problems:

1) the ambiguity, e. g. «*нашли кусок пластика*» — «*прекрасная пластика танца*» (“we found a piece of plastic”—“a great plastic of dance”);
2) the reviews often have two parts—descriptive and evaluative. The words that are sentiment-bearing for descriptive part are not those for evaluative and vice versa. In (Taboada et al., 2009) the solution of the problem of the descriptive noise is proposed;
3) the author of review’s was afraid of something but his or her fear was not confirmed;
4) for many words a number of reviews containing such words exceeds several tens (and even hundreds)—for the annotators it was hard to see all reviews in such cases;
5) the morphological errors, e.g. word «*отстой*» (“*bullshit*”) is recognized as «*отстоять*» (“*to stand*”);
6) typos, e.g. «*комплимент — комплемент*» (“*compliment—complement*”).

At the third stage positive and negative labelled word lists are joined, domain-dependent and universal sentiment lexicons are formed[8].

[8] Created sentiment lexicons are available at: https://goo.gl/KRWo5X.

5. Analysis of lexicons

5.1. Description

As a result of the proposed procedure each annotator created four lexicons for each of five domains (80 lexicons altogether). The characteristics of lexicon for restaurant domain are shown in Fig. 2.

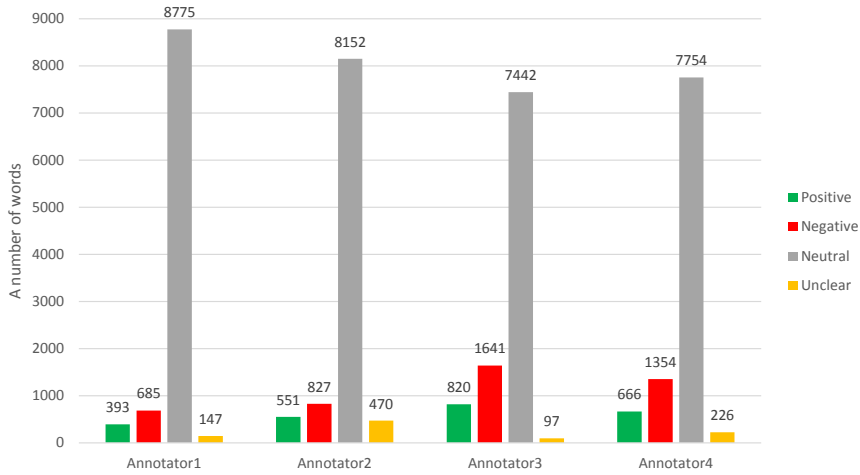


Fig. 2. The distribution of sentiment words for restaurant reviews

The analysis of created lexicons allows us to draw following conclusions. Firstly, negative lexicon is more diverse: on average the size of negative lexicons is 1.63 times more than of positive ones, despite the fact that the positive words prevail in texts (Boucher, Osgood, 1969). Secondly, the annotators differ in the degrees of confidence in their labels: the average rate of unclear words varies from 0.5% to 3.6%. At the same time the intersections of all or the most part of manual lexicons give good results of sentiment analysis comparable with automatic dictionaries (see Section 6).

Thirdly, the rate of sentiment lexicon ranges from 8.4% to 17.3% on average for various domains (Table 3). It should be noted that this rate is in specially collected dictionary of candidate words. For the full dictionary this rate is likely to be even lower.

Table 3. Average sizes of lexicons

Domain	Positive	Negative	Neutral	Unclear	Total	(Pos+Neg)/Total
Restaurants	608	1,127	8,031	235	10,000	17.3%
Cars	429	975	8,444	152	10,000	14.0%
Movies	389	451	9,026	134	10,000	8.4%
Books	491	623	8,754	132	10,000	11.1%
Cameras	535	965	8,382	119	10,000	15.0%

5.2. Intersections and unions

We built the intersection of two types of lexicons: for which all 4 annotators agree and for which at least 3 of 4 annotators agree. The characteristics of these lexicons are shown in Tables 4 and 5.

Table 4. Lexicons with 4 agreed annotators

Domain	Positive	Negative	Neutral	Unclear	Total	Part of labelled dictionary
Restaurants	200	410	6,673	0	7,283	72.8%
Cars	87	159	7,183	0	7,429	74.3%
Movies	87	109	8,123	0	8,319	83.2%
Books	109	155	7,786	1	8,051	80.5%
Cameras	79	89	6,969	0	7,137	71.4%
Average	112	184	7,347	0	7,644	76.4%

Table 5. Lexicons with the minimum 3 agreed annotators

Domain	Positive	Negative	Neutral	Unclear	Total	Part of labelled dictionary
Restaurants	483	857	7,740	14	9,094	90.9%
Cars	342	780	8,091	2	9,215	92.2%
Movies	251	317	8,873	2	9,443	94.4%
Books	359	477	8,507	1	9,344	93.4%
Cameras	396	739	7,974	3	9,112	91.1%
Average	366	634	8,237	4	9,242	92.4%

The study of Tables 4 and 5 shows the decrease in scattering of labelled lexicons parts in the transition from the agreement of all annotators to an agreement of at least three of them: from [71.4%...83.2%] to [90.9%...94.4%]. Thus, the degree of agreement of the majority is higher than 90%.

Also the universal dictionaries were created—the unions of dictionaries for all domains with different minimum number of agreed annotators (Table 6).

Table 6. The characteristics of universal lexicons

A minimum number of agreed annotators	Positive	Negative	Positive ∪ Negative	Neutral	Unclear
1	2,731	4,978	7,526	25,688	2,324
2	1,614	3,338	4,927	24,260	260
3	1,047	2,210	3,247	23,026	22
4	388	724	1,111	21,145	1

It may be noticed that the size of positive and negative lexicons union is less than the sum of positive and negative lexicons sizes separately. The reason is that some words occur in positive and negative lexicons simultaneously. For example in Table 7 there are 10 such words for the minimum three agreed annotators.

Table 7. Words belonging to both universal lexicons

Word	Positive lexicon		Negative lexicon	
	Domain	**Examples**	**Domain**	**Examples**
засасывать	books	*сюжет засасывает*	cameras	*засасывает пыль*
предсказуемость	cars	*предсказуемость в поворотах*	movies, books	*предсказуемость интриги*
непредсказуемость	books	*сюжет нравится непредсказуемостью*	cars, cameras	*непредсказуемость результата съемки*
предсказуемый	cars, cameras	*предсказуемо ведет себя*	books	*конец предсказуем*
непредсказуемый	movies, books	*непредсказуемые реакции героев*	cars, cameras	*непредсказуемые отказы*
простенько	cameras	*все простенько и со вкусом*	books	*слишком простенько*
цеплять	books	*книга цепляет за живое*	cars	*цепляет днищем землю*
затрепывать	books	*книга уже затрепана*	restaurants	*инвентарь затрепан*
реветь	books	*ревела в три ручья*	cars	*мотор ревет*
разжевывать	cameras	*разжевано для «тормозов»*	books	*разжеванный автором до неприличия*

5.3. Inter-annotator agreement

We compute inter-annotator agreement by means of Fleiss' kappa statistical measure (Fleiss, 1971). It is calculated as the ratio of degree of annotators agreement actually attained above what would be predicted by chance and the degree of agreement attainable above chance:

$$\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e},$$

where $\overline{P}$—the mean of the proportions of agreeing annotator-annotator pairs for each word; $\overline{P}_e$—the degree of agreement expected by chance.

If the annotators are in complete agreement then $\kappa = 1$. If there is chance agreement then $\kappa = 0$.

Also we compute inter-annotator agreement for each category—positive, negative, neutral and unclear. The results are shown in Table 8.

Table 8. Inter-annotator agreement

Domain	Positive	Negative	Neutral	Unclear	Fleiss' kappa
Restaurants	0.353	0.364	0.790	0.027	0.535
Cars	0.317	0.306	0.796	0.017	0.471
Movies	0.248	0.284	0.877	0.011	0.462
Books	0.297	0.322	0.849	0.019	0.504
Cameras	0.262	0.274	0.775	0.017	0.432
Average	0.295	0.310	0.817	0.018	0.481

The obtained values of Fleiss' kappa (from 0.432 for cameras to 0.535 for restaurants) on the scale from paper (Landis, Koch, 1977) refer to "the moderate agreement" (0.4...0.6). Although (Artstein, Poesio, 2008) indicate, that only values above 0.8 ensured an annotation of reasonable quality, our experiments show that the created lexicons are of sufficient quality for sentiment analysis (see Section 6).

The relatively low value of Fleiss' kappa = 0.432 for the cameras, is possibly due to a lesser awareness of annotators in this domain than in others.

Note that Fleiss' kappa was lower for movies regarding restaurants (despite the high degree of agreement in the Table 4), due to the high values of the degree of agreement P_e expected by chance.

5.4. Parts of speech

We analyzed parts of speech distribution in the unions of positive and negative lexicons for different domains (see Table 5), formed by at least 3 agreed annotators (Table 9).

Table 9. The distribution of parts of speech

	Nouns		Verbs		Adjectives		Adverbs		Others		Total	
Domain	#	%	#	%	#	%	#	%	#	%	#	%
Restaurants	336	25.1	276	20.6	512	38.2	215	16.0	1	0.1	1,340	100
Cars	281	25.0	338	30.1	377	33.6	125	11.1	1	0.1	1,122	100
Movies	146	25.7	72	12.7	226	39.8	121	21.3	3	0.5	568	100
Books	189	22.6	141	16.9	334	40.0	171	20.5	1	0.1	836	100
Cameras	255	22.5	294	25.9	437	38.5	148	13.0	1	0.1	1,135	100
Universal	865	26.6	834	25.7	1,118	34.4	428	13.2	2	0.1	3,247	100
Average	241	24.6	224	22.0	377	37.4	156	15.9	1.5	0.2		

As a result of the analysis it was found that adjectives occupy the largest part in the sentiment dictionaries (on average 37.4%). Adverbs have the smallest part (15.9%), except for *Others*. Nouns and verbs have approximately the same proportion (24.6% and 22%, respectively).

Verbs have the highest variation of proportions in the domains: from 12.7% for movies to 30.1% for cars. This is probably due to the predominance of actions description in the reviews of the goods (cameras, cars), than in the reviews of the works of art (movies, books).

5.5. Interconnection between manual and automatic lexicons

We compared the sentiment lexicons created by annotators (minimum three agreed) and automatically generated based on the weight RF. If the size of manual lexicon is equal to *N*, we take *N* first words with maximal RF-weights (Table 10).

You may notice that in general, the coincidence is low—on average 17.1% in all lexicons and domains. At the same time the scattering is very large: for the positive—from 6.0% to 33.3%, for the negative—from 11.0% to 31.3%.

Therefore, you should not rely only on automatic methods for sentiment lexicon creating. For example, top-100 RF-weighted positive words for the books domain contains such neutral words as *подход (an approach)*, *окружающий (surrounding)*, *сестра (a sister), вставать (to stand up)*, *держать (to hold)*, etc. In our opinion, the used hybrid approach where human annotators mark up a subset of the words selected by automatic methods is more promising.

Table 10. A comparison of manual and automatic lexicons

Domain	Positive		Negative	
	Size	Coincidence	Size	Coincidence
Restaurants	483	33.3%	857	31,3%
Cars	342	15.5%	780	20.1%
Movies	251	6.0%	317	11.0%
Books	359	19.2%	477	19.1%
Cameras	396	15.9%	739	14.2%
Average	366	18.0%	634	16.1%

6. Comparison of lexicons in automatic sentiment analysis

We researched the performance of the sentiment analysis for different domains using prepared sentiment lexicons and compared with the dictionaries automatically formed on the basis of train collections, as well as with the existing lexicons (see Section 2).

A vector space model of text representation was used. Automatically created dictionaries based on the training collection were weighted using an RF scheme (Lan et al, 2009). Also a feature selection was applied for the dictionaries—the first *p*% of words with the highest weights were selected. The ratio *p* ranged from 10% to 100% with 10% step. For the other dictionaries the binary weights were used.

The method SVM from scikit-learn package (Pedregosa et al., 2011) was used for classification. The kernel (linear, polynomial, RBF), SVM parameters and parameter p in the feature selection through grid search and 3-fold cross-validation were selected. The best results were achieved with a linear kernel.

We included in testing the formed by annotators domain-dependent sentiment lexicons, which contained only the words about which agree all annotators (denoted "Domain, $n = 4$") and most annotators ("Domain, $n = 3$"). In addition, we used universal sentiment lexicons ($n = 3$ and $n = 4$).

We also compared the quality of the analysis with the results of publicly available Russian sentiment lexicons: ProductSentiRus (Chetviorkin, Loukachevitch, 2012), NRC (Mohammad, Turney, 2013) and Chen-Skiena (Chen, Skiena, 2014). The sizes of all the lexicons are listed in Table 11.

As a baseline we used dummy classifier, which categorized all the objects as positive.

For evaluation we used F1-measure, for which macro-averaging was carried out due to the strong imbalance of test corpora. The test results are shown in Table 12.

Table 11. Size of lexicons (for the lexicons of train collection in the brackets it shows the part p of the full lexicon—the result of feature selection)

Lexicon	Restaurants	Cars	Movies	Books	Cameras
Dictionaries of train corpus (RF)	21,454 (1.0)	12,467 (0.7)	23,164 (0.8)	15,328 (1.0)	9,781 (0.7)
Domain ($n = 4$)	610	246	196	264	168
Domain ($n = 3$)	1,340	1,122	568	836	1,135
Universal ($n = 4$)	1,111				
Universal ($n = 3$)	3,247				
ProductSentiRus	5,000				
NRC	4,590				
Chen-Skiena	2,914				

Table 12. The results of experiments—F1-measure, %

Lexicon	Restaurants	Cars	Movies	Books	Cameras	Average F1
Baseline	47.6	49.2	45.2	47.4	47.9	47.5
Dictionaries of train corpus (RF)	74.4	63.6	64.4	61.2	80.2	68.8
Domain ($n = 4$)	74.9	62.3	65.2	**64.0**	76.0	68.5
Domain ($n = 3$)	75.0	65.2	62.0	60.5	73.9	67.3
Universal ($n = 4$)	74.3	63.3	61.4	63.1	76.8	67.8
Universal ($n = 3$)	75.3	**65.8**	**65.7**	60.2	78.9	**69.2**
ProductSentiRus	**76.2**	63.6	61.7	59.2	82.6	68.7
NRC	71.8	62.2	58.6	53.6	**82.9**	65.8
Chen-Skiena	71.2	59.6	58.7	56.6	80.2	65.2

From Table 12 it can be seen that the created sentiment lexicons allow to perform the sentiment analysis with high quality, comparable or superior the auto-generated dictionaries. At the same time the size of manual lexicons is much smaller than of automatic lexicons: for example, lexicon *books* (Domain, $n = 4$) comprises a total of 264 words and shows the quality that surpasses all other lexicons (64%).

Also the universal lexicons demonstrate the high quality, for example, the universal lexicon ($n = 3$) shows the best results in two areas of the five (cars and movies), as well as on the average.

Due to the high degree of imbalance of corpora (see Table 1) and the use of macro-averaging scheme, the quality of the analysis highly depends on the F1-measure for negative texts. Almost all relatively low results in Table 12 (e.g., Chen-Skiena for cars, dictionary of train corpus for books, NRC for movies) are closely related with poor recognition of negative texts. Low results of manual lexicons for cameras also depend on it. The reason is the insufficient size of the negative lexicon for cameras (89 words, $n = 4$). Perhaps it was the result of poor awareness of annotators in a given domain.

We note also that ProductSentiRus performed well in the analysis of product reviews (cars and cameras), as well as the restaurants. Lexicons, received by automatic translation into Russian (NRC and Chen-Skiena) tend to show relatively low quality (except cameras).

7. Conclusion

Thus, the proposed in the article procedure allows creating a compact and domain-dependent sentiment lexicon, which is very effective in sentiment analysis. The laboriousness of lexicon creation is reduced through the use of automated methods of terms weighting to generate a set of words to labeling process. It is also important for annotators to be familiar with the domain.

The universal lexicons created by union of manual lexicons also show good results comparable or superior to automatic dictionaries.

We see the following directions for future research: to expand the set of domains (news, social networks, policy) to increase the reliability of research; to investigate the influence of collocations and parts of speech on the effectiveness of lexicons; to test the lexicons with lexical-based method of sentiment analysis (Taboada et al., 2011).

Acknowledgements

The reported study was funded by RFBR according to the research project No. 16-07-00342 a.

References

1. *Amiri F., Scerri S., Khodashahi M.* (2015), Lexicon-based Sentiment Analysis for Persian Text, Proceedings of Recent Advances in Natural Language Processing, Hissar, pp. 9–16.
2. *Andreevskaia A., Bergler S.* (2006), Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses, Proceedings EACL-06, the 11rd Conference of the European Chapter of the Association for Computational Linguistics, Trento, pp. 209–216.
3. *Artstein R., Poesio M.* (2008), Inter-Coder Agreement for Computational Linguistics, Computational Linguistics, Vol. 34, No. 4, pp. 555–596.
4. *Blinov P., Kotelnikov E.* (2014), Using Distributed Representations for Aspect-Based Sentiment Analysis, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2014", Bekasovo, No. 13 (20), Vol. 2., pp. 68–79.
5. *Boucher J. D., Osgood Ch. E.* (1969), The Pollyanna Hypothesis, Journal of Verbal Learning and Verbal Behavior, No. 8, pp. 1–8.
6. *Chen Y., Skiena S.* (2014), Building Sentiment Lexicons for All Major Languages, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, pp. 383–389.
7. *Chetviorkin I., Braslavskiy P., Loukachevitch N.* (2012), Sentiment Analysis Track at ROMIP 2011, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog2012", No. 11 (18), Vol. 2., pp. 1–14.
8. *Chetviorkin I., Loukachevitch N.* (2012), Extraction of Russian Sentiment Lexicon for Product Meta-Domain, Proceedings of COLING 2012, Mumbai, pp. 593–610.
9. *Chetviorkin I., Loukachevitch N.* (2013), Sentiment Analysis Track at ROMIP 2012, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog2013", No. 12 (19), Vol. 2, pp. 40–50.
10. *Feldman R.* (2013), Techniques and Applications for Sentiment Analysis, Communications of ACM, Vol. 56, No. 4, pp. 82–89.
11. *Fellbaum C.* (1998), WordNet: An Electronic Lexical Database, Cambridge, MA: MIT Press.
12. *Fleiss J. L.* (1971), Measuring nominal scale agreement among many raters, Psychological Bulletin, Vol. 76, No. 5, pp. 378–382.
13. *Freitas L., Vieira R.* (2013), Comparing Portuguese Opinion Lexicons in Feature-Based Sentiment Analysis, International Journal of Computational Linguistics and Applications, Vol. 4 (1), pp. 147–158.
14. *Ivanov V., Tutubalina E., Mingazov N., Alimova I.* (2015), Extracting Aspects, Sentiment and Categories of Aspects in User Reviews about Restaurants and Cars, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2015", Moscow, pp. 22–33.
15. *Kiselev Y., Braslavski P., Menshikov I., Mukhin M., Krizhanovskaya N.* (2015), Russian Lexicographic Landscape: a Tale of 12 Dictionaries, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog2015", No. 14 (21), Vol. 1, pp. 254–50271

16. *Lan M., Tan C. L., Su J., Lu Y.* (2009), Supervised and Traditional Term Weighting Methods for Automatic Text Categorization, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 31 (4), pp. 721–735.
17. *Landis J. R., Koch G. G.* (1977), The Measurement of Observer Agreement for Categorical Data, Biometrics, Vol. 33, pp. 159–174.
18. *Liu B.* (2012), Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies, Vol. 5 (1).
19. *Mohammad S., Turney P.* (2013), Crowdsourcing a Word-Emotion Association Lexicon, Computational Intelligence, Vol. 29 (3), pp. 436–465.
20. *Musto C., Semeraro G., Polignano M.* (2014), A Comparison of Lexicon-based Approaches for Sentiment Analysis of Microblog Posts, DART 2014 8th International Workshop on Information Filtering and Retrieval, Pisa.
21. *Ozdemir C., Bergler S.* (2015), A Comparative Study of Different Sentiment Lexica for Sentiment Analysis of Tweets, Proceedings of Recent Advances in Natural Language Processing, Hissar, pp. 488–496.
22. *Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É.* (2011), Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, Vol. 12, pp. 2825–2830.
23. *Taboada M., Brooke J., Stede M.* (2009), Genre-based Paragraph Classification for Sentiment Analysis, Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue, London, pp. 62–70.
24. *Taboada M., Brooke J., Tofiloski M., Voll K., Stede M.* (2011), Lexicon-Based Methods for Sentiment Analysis, Computational Linguistics, Vol. 37 (2), pp. 267–307.
25. *Ulanov A., Sapozhnikov G.* (2013), Context-Dependent Opinion Lexicon Translation with the Use of a Parallel Corpus, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2013”, Bekasovo, pp. 165–174.

ПРОБЛЕМА ИДЕНТИФИКАЦИИ ДИКТОРА В УСЛОВИЯХ ШЕПОТНОЙ РЕЧИ

Крейчи С. А. (kreychi@mail.ru),
Кривнова О. Ф. (okrivnova@mail.ru),
Ступина Е. А. (ek.stupina@gmail.com)

МГУ им. М. В. Ломоносова, Москва, Россия

Ключевые слова: диктор, аудитор, речевой портрет, вербальный, паравербальные, экстравербальный, идентификация, артикуляция, индивидуальный, признак, нейтральная фонация, шепотная речь

THE PROBLEM OF SPEAKER IDENTIFICATION IN WHISPERED SPEECH

Kreychi S. A. (kreychi@mail.ru),
Krivnova O. F. (okrivnova@mail.ru),
Stupina E. A. (ek.stupina@gmail.com)

Moscow State Lomonosov University, Moscow, Russia

The problem of speaker identification in whispered speech is of some interest for cognitive science, as well as for forensic and language minority. The work is devoted to an experimental study of the problem of familiar and unfamiliar speaker identification in vocal speech and in whisper. The experiment simulated in some respects the task and conditions of identification of a speaker by a listener. The initial number of participants—18 persons has been increased with the help of a questionnaire survey on the Internet. The number of remote auditors amounted to 125 people. The experiment took place in "online". The subject listened to and identified at least 16 pairs of entries. At the end of the experiment, he indicated if he had recognized any of the speakers. It was found that the main clues for correct recognition of a familiar speaker are the individual characteristics of his articulation, the components of the extraverbal part of his "speech portrait". The other features, such as individual speech style and individual manner of pauses and text macrosegmentation did not have any significant effects on speaker identification in whispered speech.

Key words: speaker, listener, speech portrait, verbal, extraverbal, identification, neutral phonation, whisper

1. Введение

При достаточно длительном периоде речевой коммуникации в памяти человека формируется «речевой портрет» собеседника, включающий характеристики разного рода: вербальные (остаточные диалектные или иноязычные явления и т. п.), паравербальные (индивидуальный способ передачи эмоциональных состояний) и экстравербальные составляющие (Потапов, Потапова 2008): типичный для говорящего темп и громкость речи, характерный способ макросегментации речевого потока и его паузирования, напряженность речевого тракта, характер коартикуляции, динамика основного тона. Экстравербальная составляющая включает в себя всю совокупность акустических параметров голоса и речи человека. Существенную роль при этом играют индивидуальные признаки целевых артикуляций, соответствующих фонемам и их звуковым реализациям в речи (Кедрова, Захаров, Анисимов, Пирогов 2003); (Kedrova, Borisoff 2013). Им соответствуют в памяти человека слуховые образы, которые также входят в «речевой портрет» говорящего. Механизм и навыки артикуляции формируются у человека с детства по мере обучения языку и устной речи, в совокупности с определенными индивидуальными признаками, которые сохраняются и в условиях шепотной речи, когда отсутствует нормальная голосовая фонация и дополнительная информация о тембральных характеристиках голоса и просодическом оформлении речевого сообщения. Тем не менее, повседневный коммуникативный опыт показывает, что знакомый собеседник, как правило, узнается по остальным параметрам речи, в том числе и за счет избыточности информации в речевом сигнале.

Вопросы идентификации личности говорящего по шепотной речи представляют определенный интерес для когнитивной науки, а также для форенсики и лингвокриминалистики, так как шепотная речь является частью доказательной базы при фоноскопическом анализе зафиксированных разговоров фигурантов уголовного дела. При этом в качестве образцов речи собеседников эксперту-фоноскописту предоставляются только фонограммы с использованием голоса.

Исследование формантной структуры звуков речи, произнесенных с участием голоса и шепотом (Костина 2013), показало, что резонаторы речевого тракта реагируют одинаково как на воздушный поток, модулированный импульсами голосовых связок, так и на турбулентный поток, образованный при прохождении воздуха через межхрящевую щель при замкнутой полностью или частично междусвязочной щели (Зиндер 1979), (Кодзасов, Кривнова 2001). По полученным данным спектральная картина шепотной речи сохраняет основные черты спектральной картины того же высказывания, произнесенного тем же диктором, но с участием голоса. Исследование Костиной выявило также некоторые особенности шепотной речи. Так, более напряженная работа артикуляторов и более высокая степень произносительных усилий приводит к увеличению общего времени произнесения и длительности пауз. Среднезвуковая длительность в шёпотной речи в среднем на 5 % больше, чем в речи с участием голоса. Значительно (на 20 %) увеличивается общее время пауз. Это, видимо, обусловлено бóльшим расходом воздуха при шепоте

и увеличением количества дыхательных пауз. В шепотной речи присутствуют также определенные корреляты частоты основного тона голоса (ЧОТ) (Лиханова 2007). Так, на гласных, примерно в той области, где в обычной речи происходит изменение ЧОТ, в шёпотной речи наблюдаются изменения частоты первых спектральных пиков.

Наша работа посвящена экспериментальному исследованию идентификации знакомого и незнакомого диктора по речи с голосом и в условиях шепотного произнесения. Эксперимент имитировал в определенной степени задачу и условия идентификации личности говорящего слушающим. В ходе эксперимента планировалось выявить перцептивно похожие между собой образцы речи (голоса́ в широком смысле слова), а также оценить степень сложности процедуры идентификации по шёпотной речи диктора, знакомого или незнакомого слушающему. Кроме того, планировалось также сравнить успешность идентификации говорящего при нейтральной фонации и в шёпотной речи.

Объектом исследования была экстравербальная составляющая «речевого портрета» говорящего. С целью уменьшения влияния вербальной и паравербальной частей речевого портрета на процесс идентификации, в качестве тестового материала использовались фрагменты двух текстов фактологического содержания. Для озвучивания текстов были отобраны две группы дикторов. Первая группа состояла из трех мужчин и четырех женщин, а вторая — из семи мужчин и девяти женщин. Запись звукового материала производилась на студийном оборудовании высокого класса, в дикторской кабине, изолированной от посторонних звуков. Каждый из дикторов прочитывал тексты дважды, с использованием голоса и шепотом. Дикторы принадлежали к одинаковой социальной группе, имели одинаковый уровень образования, владели хорошей техникой чтения и произносительными нормами современного русского литературного языка.

Для исследования проблемы идентификации говорящего в условиях шёпотной речи было проведено два перцептивных эксперимента. В обоих экспериментах изучалась возможность и сложность выполнения процедуры идентификации диктора по шёпотной речи. При этом первый эксперимент был подготовительным. В нем имитировалась задача и условия идентификации *знакомого* голоса и исследовались особенности идентификации голосов женщин и мужчин. Второй эксперимент был проведен с помощью метода дистанционного анкетирования. Этот метод представляется удобным для проведения тестовых экспериментов из-за простоты и возможности получить значительную по объему выборку за минимальное время. Желательно при этом, чтобы эксперимент занимал не слишком много времени. Это уменьшит количество отказов (когда участник прерывает эксперимент, не закончив его). Если эксперимент не будет утомительным, можно ожидать, что участники эксперимента не отвлекаются и концентрируют свое внимание на задании. Кроме того, предлагаемое задание должно быть простым для объяснения. Непосредственно перед выполнением основного задания следует провести краткую тренировку, чтобы участник ознакомился с форматом задания.

Однако данный формат тестирования имеет ряд ограничений, связанных с отсутствием прямого контакта испытуемого и экспериментатора.

Экспериментатор не имеет возможности следить за ходом эксперимента. Даже если само задание является простым и недолгим, у экспериментатора все равно нет возможности контролировать внимание участника во время эксперимента. Еще один немаловажный фактор, который можно контролировать с помощью опросника, — это разнообразие технических средств, с помощью которых происходит прослушивание экспериментальных стимулов. Кроме того, обычно подобные опросы распространяются через знакомых или социальные сети, что может привести к неслучайной выборке (т. н. snowball sampling).

В исследованиях восприятия речи используются, в основном, эксперименты двух типов: на идентификацию и на различение стимулов (identification and discrimination (Pols Louis 1989). Выбранный нами экспериментальный дизайн (discrimination task: same — different) представляется довольно удобным для дистанционного режима. Задание является простым для объяснения, участникам не требуется разъяснять, на какого рода особые различия в стимульном материале им следует обращать внимание. Им необходимо лишь отметить, одному ли диктору принадлежат оба прослушиваемых речевых отрезка. Благодаря тому, что участники принимают решение только после прослушивания второго стимула в паре, время реакции является достаточно надежным оценочным показателем, и его легко измерить. Одним из недостатков данного дизайна является, однако, возможная тенденция аудиторов отвечать «один и тот же», особенно, если задание кажется им сложным. В таком случае количество ошибочных ответов «один и тот же» становится неоправданно больше, чем количество ошибочных ответов «разные».

Наш эксперимент сохраняет высокую экологическую валидность ввиду естественности стимулов (обработка материала была минимальной: только нормализация по громкости). Для того, чтобы максимально приблизить речевой материал к естественному, мы использовали фрагменты прочитанных фактологических текстов, при этом сохранив идентичный лексический состав материала по всем дикторам. Таким образом, исследовалась возможность и сложность идентификации дикторов именно по экстравербальной информации.

2. Дизайн, задачи и результаты первого перцептивного эксперимента I

Для эксперимента I было выбрано 9 аудиторов, хорошо знакомых с дикторами первой группы (7 человек), и не знакомых с дикторами второй группы (16 человек).

Тестовый материал в данном эксперименте прослушивался с целью идентификации знакомого диктора в шепотной речи. Тембрально-высотные характеристики голосовой речи диктора являются дополнительной информацией, откладывающейся в памяти слушателя. Для исключения влияния речи с голосом на идентификацию того же диктора по шепоту каждому испытуемому предъявлялись в случайном порядке звуковые записи сначала все, произнесенные шепотом, а потом все с голосом. Испытуемый слушал очередную запись и, определившись

с ответом, произносил имя знакомого диктора. При этом ответ аудитора и время от начала звучания текстового фрагмента до принятия решения фиксировались оператором. Каждый ответ оценивался как «+», если знакомый диктор был идентифицирован правильно, и «–» если неправильно. В таблицах 1–4 представлены результаты прослушивания. Рядом с результатами правильной идентификации указано время принятия решения в секундах (в скобках).

Таблица 1. Результаты прослушивания текстов, прочитанных женскими голосами с нейтральной фонацией. В строках содержатся данные по каждому диктору f1–f4, а в столбцах — результаты и время идентификации (в секундах) аудиторами 1–9

Дикторы	Аудиторы								
	1	2	3	4	5	6	7	8	9
f1	+ (2)	–	+ (3)	–	+ (3)	+ (3)	+ (4)	–	+ (7)
f2	+ (12)	+ (13)	+ (29)	+ (6)	–	+ (1)	+ (3)	–	–
f3	–	–	+ (2)	+ (12)	–	+ (13)	+ (3)	–	+ (6)
f4	+ (3)	+ (2)	+ (1)	+ (2)	+ (2)	+ (2)	+ (2)	+ (1)	+ (2)

Таблица 2. Результаты прослушивания текстов, прочитанных женскими голосами шепотом. В строках содержатся данные по каждому диктору f1–f4, а в столбцах — результаты и время идентификации (в секундах) аудиторами 1–9

Дикторы	Аудиторы								
	1	2	3	4	5	6	7	8	9
f1	+ (15)	–	+ (9)	–	+ (10)	–	+ (13)	–	+ (11)
f2	–	–	–	+ (23)	–	–	+ (3)	–	–
f3	–	–	+ (13)	–	–	+ (35)	+ (4)	–	+ (26)
f4	+ (2)	+ (3)	+ (2)	+ (2)	+ (5)	–	+ (2)	+ (1)	+ (1)

Таблица 3. Результаты прослушивания текстов, прочитанных мужскими голосами с нейтральной фонацией. В строках содержатся данные по каждому диктору m1–m3, а в столбцах — результаты и время идентификации (в секундах) аудиторами 1–9

Дикторы	Аудиторы								
	1	2	3	4	5	6	7	8	9
m1	+ (2)	+ (2)	+ (2)	+ (3)	+ (1)	+ (1)	+ (4)	+ (1)	+ (9)
m2	+ (1)	+ (1)	+ (1)	+ (1)	+ (1)	+ (1)	+ (1)	+ (1)	+ (1)
m3	+ (3)	+ (2)	+ (1)	+ (2)	+ (2)	+ (1)	+ (2)	+ (2)	+ (1)

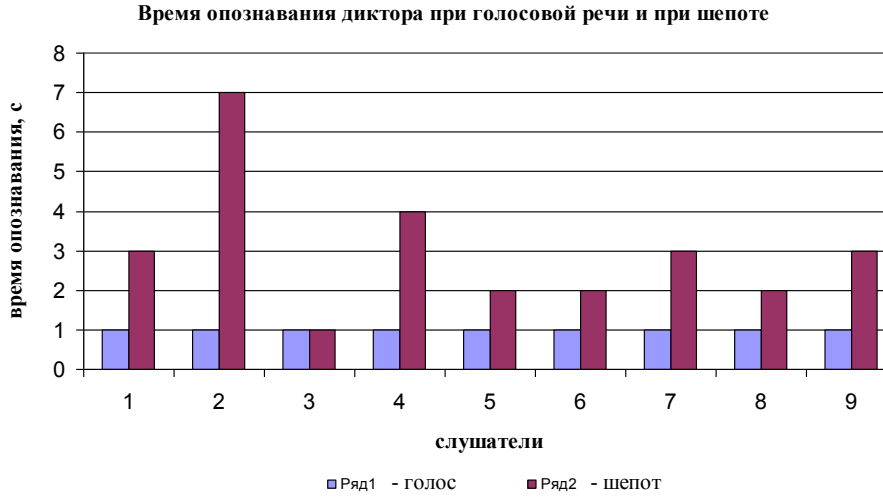
Таблица 4. Результаты прослушивания текстов, прочитанных мужскими голосами шепотом. В строках содержатся данные по каждому диктору m1–m3, а в столбцах — результаты и время идентификации (в секундах) аудиторами 1–9

Дикторы	Аудиторы								
	1	2	3	4	5	6	7	8	9
m1	+ (6)	+ (15)	+ (3)	+ (20)	+ (5)	+ (4)	+ (10)	+ (3)	–
m2	+ (3)	+ (7)	+ (1)	+ (4)	+ (2)	+ (2)	+ (3)	+ (2)	+ (3)
m3	–	+ (4)	+ (1)	+ (3)	+ (4)	+ (4)	+ (3)	+ (2)	+ (4)

Из таблиц 1–4 видно, что задача опознания знакомого женского голоса была выполнена аудиторами правильно на 72,2%, а опознания того же диктора по шепоту — на 52,8%. Для мужских голосов задача опознания знакомого диктора была выполнена правильно на 96,3%, а по шепоту — на 92,6%. Ниже приведена диаграмма, иллюстрирующая сравнительное *время* опознавания стопроцентно узнаваемого диктора при голосовой речи и при шепоте.

Заметна высокая степень узнаваемости диктора f4, носившего брекеты, что влияло на артикуляцию и создавало дополнительные индивидуальные экстравербальные признаки. Впрочем, этого можно было ожидать и заранее. Кроме того, наблюдались случаи отсутствия узнавания аудитором собственного голоса. Это объясняется тем, что при восприятии собственной речи звуковые волны приходят к органам слуха человека, в основном, не через внешнее пространство, а через ткани головы. При этом акустические параметры речевых сигналов отличаются. Поэтому для узнавания себя в звукозаписи требуется некоторый опыт прослушивания.

Результаты эксперимента I показали, что мужские голоса (как в шёпотной, так и в голосовой речи) идентифицируются аудиторами более успешно. Кроме того, наблюдаемая разница между результатами двух серий эксперимента (шёпот, голос) для мужских и женских голосов, различна. Для мужских голосов идентификация по шёпоту была хуже идентификации по речи с голосом всего на 3,7%, в то время как для женских — на 19,4%. При этом на узнавание мужских голосов аудиторам требовалось значительно меньше времени. Для мужчин среднее время узнавания диктора при нейтральной фонации равно 1,6 сек, а в условиях шёпота — 4,78; аналогичные показатели для женской речи таковы: среднее время узнавания по голосу — 5,9 сек, а по шёпоту — 14,46 сек. Эти результаты подтверждают известные наблюдения о том, что мужские голоса более ярко отражают индивидуальные особенности речеобразования, вследствие чего легче поддаются идентификации (Златоустова 1986), (Штерн 1984).



Разница во времени идентификации разных дикторов, обнаруженная в эксперименте I, послужила основанием для проведения второго перцептивного эксперимента с целью выявления наиболее значимых особенностей речи для восприятия и идентификации диктора по шепоту.

3. Дизайн, задачи и результаты второго перцептивного эксперимента II

Основной целью эксперимента II являлась проверка возможности объединения мужских голосов в группы по перцептивному сходству при шёпоте, а также выделение ключевых признаков шёпотной речи, релевантных для идентификации дикторов. В эксперименте использовались те же записи, что и в эксперименте I, но были взяты только небольшие фрагменты, прочитанные десятью дикторами мужского пола в возрасте от 16 до 32 лет. По длительности фрагменты были трех типов: 1-й тип (длинные) — 27 слов, 3 предложения, общая длительность около 10 секунд; 2-й тип (короткие) — 6 слов, 1 предложение, длительность около 3 секунд; 3-й тип (средние) — 14 слов, 1 предложение длительность около 6 секунд.

С целью определения опорных признаков для идентификации диктора, был проведен аудитивный анализ записей, использованных в эксперименте I, при котором у отдельных дикторов были выявлены характерные экстравербальные признаки речи, воспринимаемые как индивидуальные (картавость звука [r], «скрипучий» тембр голоса), индивидуальная манера интонирования тестовых фраз, паузирования, а также индивидуальная манера макросегментации длинного фрагмента из трех предложений. Результаты измерения длительностей отрезков фонации и пауз у десяти дикторов представлены в таблице 5.

Таблица 5. Длительность фрагмента из трех предложений и длительность пауз между синтагмами в шёпотной и обычной речи, сек. Строки таблицы соответствуют номерам дикторов, а столбцы — измеренным длительностям. Прочерк означает отсутствие паузы

Диктор	Общее время звучания		Длительность паузы между синтагмами 1 и 2		Длительность паузы между синтагмами 2 и 3	
	голос	шепот	голос	шепот	голос	шепот
1	10,1	9,9	0,75	0,58	0,85	0,75
2	8,9	8,6	0,43	0,70	0,79	0,5
3	8,8	9,1	0,34	0,39	0,56	0,6
4	9,5	9,8	—	—	0,14	-
5	8,3	9,5	0,54	0,70	0,75	0,69
6	8,4	9,3	0,61	0,35	0,50	0,57
7	8,6	9,6	0,53	0,18	0,09	0,35
8	8,9	9,7	0,14	0,56	0,36	0,59
9	9,3	10,1	0,30	1,12	0,38	0,66
10	10,2	10,5	0,60	1,03	0,79	0,81

Самый медленный темп речи на данном отрезке в три предложения показывает диктор 10 — самое большое общее время фонации и общая длительность пауз. Этот показатель может оказаться опорным признаком для идентификации данного диктора при сравнении образцов голоса и шёпота. Диктор 10 также показал самый медленный темп речи как в голосе, так и в шёпоте на записи типа 3, а диктор 4, имея довольно большое общее время фонации, вообще не делал пауз между предложениями. Манера чтения у этого диктора достаточно монотонная, что также может стать хорошим признаком для его идентификации.

Аудитивный анализ позволил выделить несколько возможных опорных признаков для выполнения процедуры идентификации дикторов-мужчин. На основании проведенного анализа можно было предположить, что некоторые дикторы — диктор 6 (имеющий дефект речи — картавость), диктор 10 (имеющий медленный, размеренный темп речи), диктор 4, который отличается монотонной манерой чтения без выраженных пауз, диктор 2, который имеет несколько особенностей речи, в том числе заметное придыхание в конце фраз, будут лучше отождествляться испытуемыми, по сравнению с другими дикторами.

На начальном этапе эксперимента II использовалась только одна переменная — личность диктора в рамках одной фонационной модальности. Каждому аудитору парами предъявлялись идентичные записи одного из вышеприведенных текстов, произнесенного одним и тем же диктором либо голосом, либо шёпотом. При этом типы записей по длительности в предъявляемой паре совпадали. Аудитор должен был определить, одному человеку или разным людям принадлежат обе записи в паре. В эксперименте приняло участие 18 аудиторов

(9 мужчин и 9 женщин) в возрасте от 21 до 32 лет, с высшим или незаконченным высшим образованием, для которых русский язык являлся родным. Все испытуемые были заведомо незнакомы с дикторами и не знали их голосов. Задачей каждого аудитора было отождествление диктора в предъявляемой паре записей по образцам обычной и шёпотной речи. Каждый испытуемый должен был выполнить 10 обязательных заданий (5 заданий на шёпот и 5 заданий на обычную речь), после чего они могли либо закончить тестирование, либо продолжить (далее испытуемый мог в любой момент закончить тест). Оказалось, что многие испытуемые добровольно продолжали эксперимент и после выполнения обязательных заданий. В среднем каждый испытуемый выполнил 29 заданий. При этом двое испытуемых выполнили намного больше заданий, чем другие — 115 и 151 заданий. Всего восемнадцати испытуемым было предъявлено 529 пар записей (из них 268 пара записей шёпота и 261 пара записей голоса). В сумме испытуемые ошиблись 25 раз из 529. То есть, средний процент правильных ответов был очень высоким — 95, 27 %. Большинство ошибок, как и ожидалось, пришлось на пары записей шёпота — 21. Средний процент правильных ответов в заданиях с записями голоса составил 98,47 %, а в заданиях с записями шёпота — 92,16 %.

В дальнейшем количество участников эксперимента II было увеличено с помощью анкетного опроса в интернете. Число дистанционных аудиторов составило 125 человек (71 женщина и 54 мужчины) в возрасте от 10 до 67 лет, не принимавших участие в предварительном эксперименте, для всех русский язык являлся родным. Эксперимент проходил в режиме «онлайн» на сайте http://poll.freshswag.ru/. Испытуемый заполнял анкету (опросный лист, содержавший данные о возрасте, поле, образовании и родном языке). Далее следовало задание: «*Прослушайте две записи. Постарайтесь определить, принадлежат ли данные аудиозаписи одному и тому же человеку. Вы можете прослушивать каждую запись столько раз, сколько вам потребуется для получения устойчивого решения. Своё решение зафиксируйте*».

Испытуемый прослушивал и отождествлял, по крайней мере, 16 пар записей. В конце эксперимента он указывал, узнал ли он кого-либо из дикторов, а также давал комментарий по поводу того, на что он ориентировался при выполнении задания. В опросной анкете испытуемый также указывал данные об оборудовании, которое использовалось для прослушивания.

Полученные результаты показали, что все испытуемые успешно справляются с задачей отождествления дикторов в рамках одной и той же фонационной модальности. После этого в эксперимент была введена вторая переменная — модальность фонации. Аудиторам парами предъявлялись записи двух дикторов (одного и того же или разных) и при этом в разных модальностях фонации в одной паре (голос + шёпот). Типы записей по длительности в одной паре совпадали. Результаты идентификации дикторов в разных модальностях фонации в одной паре представлены в таблице 6.

Таблица 6. Средний процент правильного отождествления каждого из 10-и дикторов в разных модальностях фонации (по результатам 125 аудиторов)

Диктор	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
кол-во предъявлений	92	96	88	101	106	69	50	100	87	91
кол-во ошибок	38	49	41	42	39	15	25	42	29	30
% правильных ответов	58,7	49,0	53,4	58,4	63,2	78,3	54,5	58,0	66,7	67,0

Полученные результаты не позволяют сделать каких-либо определенных выводов относительно использования индивидуальной манеры паузирования и макросегментации текста в качестве опорных признаков для идентификации диктора в шепотной речи.

Мужские голоса как в шёпотной речи, так и при нейтральной фонации идентифицируются аудиторами лучше, чем женские. Успешность отождествления диктора в парах записей с разной фонационной модальностью практически не зависит от пола и возраста диктора. Длительность записи не влияет на процент правильных ответов. Не было выявлено также влияния общего времени и количества прослушиваний записи на успешность выполнения задания.

4. Заключение

Проведенные нами эксперименты выявили ряд нерешенных и сложных проблем. Оказалось, что задача идентификации незнакомого диктора (особенно если в распоряжении слушающего имеются лишь небольшие фрагменты его голоса и шёпота), достаточно сложна. На данном этапе проведенного нами анализа остается открытой и проблема выделения и веса признаков, опорных для идентификации диктора.

Литература

1. *Зиндер Л. Р.* (1979) Общая фонетика. Высшая школа, М.
2. *Златоустова Л. В.* (1986) Речевые тексты и их применение. М.
3. *Кедрова Г. Е., Захаров Л. М., Анисимов Н. В., Пирогов Ю. А.* (2003) Исследование артикуляторной базы русского языка методами магнитно-резонансной томографии// XIII сессия Российского акустического общества. Т. 3. Акустика речи. Медицинская и биологическая акустика, ГЕОС, М.
4. *Кодзасов С. В., Кривнова О. Ф.* (2001) Общая фонетика. РГГУ, М.
5. *Костина А. А., Крейчи С. А.* (2013) Особенности идентификации диктора по шепотной речи. //Материалы конференции «Комплексная безопасность Отечества»// Парламентский Центр, М., с. с. 156–161.

6. *Лиханова И.* (2007) Корреляты частоты основного тона в шёпотной речи. Конкурсная работа под руководством Скрелина П. А. СПбГУ, СПб.
7. *Потапов В. В., Потапова Р. К.* (2008) Перспективы развития концепции «речевой портрет говорящего» // Материалы XVII Международной научной конференции «Информатизация и информационная безопасность правоохранительных органов», М., с. с. 381–382.
8. *Штерн А. С.* (1984) Артикуляционные таблицы. Методическая разработка для развития навыков аудирования и тестирования слуховой функции. Л.

References

1. *Kedrova G. E., Zakharov L. M., Anisimov N. V., Pirogov J. A.* (2003) The study articulatory base of the Russian language using magnetic resonance imaging [Issledovanie artikulatornoj bazy russkogo jazyka metodami magnitnorezonansnoj tomografii]// XIII session of the Russian acoustical society. Acoustics of speech. Medical and biological acoustics, GEOS, M, p. p. 81–84.
2. *Kedrova G., Borissoff L.* (2013) The concept of 'basis of articulation' in Russia in the first half of the 20th century // Historiographia Linguistica. Vol. 40, no. 1/2, p. p. 151–198.
3. *Kodsasov S. V., Krivnova O. F.* (2001) General phonetics [Obshchaja fonetika], RGGU, M.
4. *Kostina A. A., Kreychi S. A.* (2013) Features of speaker identification in whisper speech. [Osobennosti identifikatsii diktora po shëpotnoj rechi] Materials of the conference "Integrated security of Fatherland". The Parliamentary Centre, M, pp. 156–161.
5. *Likhanova I.* (2007) Correlates of the frequency of the pitch in whispered speech. [Korrelaty chastoty osnovnogo tona v shëpotnoj rechi] Competitive work under the guidance Skrelin P. A., St. Petersburg University, St. Petersburg.
6. *Potapov V. V., Potapova R. K.* (2008) Prospects of development of the concept of "verbal portrait of the speaker"[Perspektivy razvitija kontseptsii "rechevoj portret govorjashchego"]. Materials of XVII International scientific conference "Informatization and information safety of law enforcement". M., p. p. 381–382.
7. *Pols Louis C. W.* (1989) Studying speech perception: from simple stationary stimuli to more and more complex speech(-like) signals. Proceedings of the 28th Acoustic Conference.
8. *Stern A. S.* (1984) Articulation tables. Methodological development for the development of listening skills and auditory function testing. [Artikulatsionnye tablitsy. Metodicheskaja razrabotka dla razvitija navykov audirovanija i testirovanija sluhovoj funktsii]. St. Petersburg.
9. *Zinder L. R.* (1979) General phonetics. [Obshchaja fonetika] Graduate school. M.
10. *Zlatoustova L. V.* (1986) Speech texts and their application. M.

ЕСТЕСТВЕННЫЙ ЯЗЫК И ЯЗЫК ГЕОМЕТРИЧЕСКИХ ЧЕРТЕЖЕЙ

Крейдлин Г. Е. (gekr@iitp.ru),
Шабат Г. Б. (george.shabat@gmail.com)

Российский государственный гуманитарный университет, Москва, Россия

В статье рассматриваются проблемы взаимодействия естественного языка и его аналога — естественно-подобного языка геометрии с языком геометрических чертежей в двух областях интеллектуальной деятельности. Это (1) синтез и анализ текстов на естественном языке и соответствующих ему невербальных знаковых кодов и (2) устная и письменная мультимодальная коммуникация. нализируются языковые и, шире, семиотические особенности рассматриваемых языков. Утверждается, что более глубокое понимание геометрических фактов и проблем достигается при одновременном хорошем владении обоими языками.

Ключевые слова: естественный язык, невербальный знаковый код, мультимодальная коммуникация, геометрия, чертёж, понимание

THE NATURAL LANGUAGE AND THE LANGUAGE OF GEOMETRIC SKETCHES

Kreydlin G. E. (gekr@iitp.ru),
Shabat G. B. (george.shabat@gmail.com)

The Russian State University for the Humanities, Moscow, Russia

The paper deals with the problems of interaction between the natural language together with its analogue—the natural-like language of geometry — and the language of geometric sketches within the two domains of intellectual activities. These are (1) synthesis and analysis of the natural language texts together with the corresponding non-verbal signs, and (2) oral and written multimodal communicative acts. Some linguistic (morphologic, syntactic and semantic) as well as semiotic peculiarities (the use of special signs, font markup, color, etc.) of the languages are discussed. The correspondence between some fragments of the natural-like language of geometry and some sketches is established. The problem of the representations of logical connectives and quantifiers in the sketches is partially solved by constructing the sketch analogs of the natural-like units and their combinations. It is stated that the more profound understanding of geometric facts and problems can be achieved by fluent knowledge of both languages and the special translation skills.

Key words: natural language, nonverbal sign code, multimodal communication, geometry, sketch, understanding

Введение

Несмотря на то, что диалог учителя с учеником в жизни любого человека играет особую роль, этот вид диалогов сравнительно мало изучен и плохо описан. Одна из причин этого, на наш взгляд, заключается в том, что во время этих диалогов происходит *переключение кодов*. А именно, наряду с естественным языком и связанными с ним невербальными знаковыми кодами, участники диалогов при необходимости переходят на специальные языки, в частности, на языки науки. Не удивительно поэтому, что лингвистику и семиотику интересуют все языки и коды, используемые в диалоге.

Описание языков науки важно не только само по себе; оно помогает также понять природу и механизмы коммуникации вообще. Изучая языки науки, исследователь вскрывает их особенности, определяет их общие и отличительные свойства в сравнении с повседневным языком. Ниже речь пойдёт об одном классе языков науки — о языках геометрии и текстах на них в письменной форме — и их использовании в преподавательской практике.

С геометрией имеют дело не только профессионалы, но и люди, от математики весьма далёкие. Это накладывает на стиль её изложения ряд ограничений и обязательств: ядерные элементы и тексты геометрии должны быть понятны не только математикам, но и более широкому кругу людей. Было бы, однако, ошибкой думать, что с развитием геометрии всегда учитывались характеристики адресатов её текстов. Фактор адресата, наряду с внутренними потребностями науки, и явился одной из причин появления разных языков геометрии.

Мы выделили пять языков геометрии: **естественно-подобный** язык, язык **чертежей**, **формальный**, **координатный** и язык **движений**. Далее речь пойдёт первых двух из них.

Настоящая статья является продолжением серии работ, написанных совместно математиком и лингвистом (см. Литературу). В центре статьи лежат природа и механизмы понимания текстов, прежде всего научных. Аналогичные вопросы рассматривались в целом ряде исследований, причём как лингвистических, так и междисциплинарных. См. Гладкий 1997; Гладкий, Крейдлин 1991; Звонкин 1990; Корельская, Падучева 1978; Крейдлин, Шмелёв 1989, 1994; Манин 1977, 2008; Падучева 1974; Пеньковский 2005.

1. Языки геометрии и их место среди языков математики

Рассматриваемые нами геометрические тексты конструируют пространство и преобразуют его. Для них характерны панхрония (независимость от времени), утвердительная модальность предложений и культурная универсальность; кроме того, в них встречаются невербальные знаки разной природы. Некоторые из таких знаков закреплены в качестве обозначений фиксированных объектов. Так, в геометрических текстах принято обозначать основные объекты латинскими буквами (точки — заглавными; прямые, окружности и т. д. — строчными), а буква π служит стандартным именем отношения длины

окружности к её диаметру. В этих текстах имеются также значки начала и конца доказательств (например, ◀▶ обрамляет тексты доказательств, а ■ ставится в конце доказательства). Существуют общепринятые (часто национально- или культурно-специфичные) соглашения[1] об использованиях определённых шрифтов (фонтов) в формулировках теорем, соглашения об аббревиатурах и др.

2. Естественно-подобный язык планиметрии

2.1. Лексика и грамматика естественно-подобного языка геометрии

Из всех языков геометрии естественно-подобный наименее формализован. Его лексику составляют определённые слова и словосочетания естественного языка, а также единицы, которые являются вокабулами и лексемами словарей математических терминов. Из слов русского языка в естественно-подобный язык геометрии попадают только те, которые хорошо сочетаются с математической лексикой. Например, в него не входят экспрессивно окрашенные слова, глаголы, выражающие коммуникативное поведение человека, жестовые фразеологизмы (Крейдлин 2002) и единицы ряда других лексических и семантических классов. С другой стороны, в лексику этого языка входят как всем известные со школы слова и словосочетания *катет*, *медиана*, *параллелограмм*, *вписать окружность* и т. п., так и единицы *поляра*, *чевиана*, *инвертировать* и пр., не входящие в стандартный школьный язык.

Можно отметить также особенность сочетаемости некоторых слов, омонимичных лексемам бытового русского языка или полисемичных им. Так, можно сказать ***опустить высоту*** (даже на боковую сторону треугольника!), но не **опустить медиану*. У равнобедренного треугольника выделяют *основание* и *боковые* стороны, причём, вопреки сложившемуся словоупотреблению, основание может лежать не внизу чертежа. Существует словосочетание ***вертикальные углы***, но нет сочетания **вертикальный угол*.

Словообразовательный потенциал геометрических терминов, вообще говоря, уже, чем у слов той же части речи бытового языка. Так, не образуются прилагательные от существительных *катет*, *гипотенуза*, *окружность*; нет кратких форм у прилагательных *равносторонний*, *квадратный*; редки сложные слова и общепринятые аббревиатуры (одно из немногих исключений — аббревиатура ГМТ, означающая Геометрическое Место Точек).

Переходя к характеристике грамматики, начнём с морфологии. Она достаточно бедная: нет, например, морфологических показателей категорий уменьшительности и ласкательности: нельзя сказать **гипотенузка* или **равнобедренненький*. Основные содержательные типы фраз этого языка панхроничны,

[1] По всей видимости, в явном виде нигде не сформулированные.

а потому глаголы в них не содержат временных маркеров, таких, как суффикс прошедшего времени *-л*; невозможны в них и аналитические временные формы вроде **были перпендикулярны*.

Что касается синтаксиса, то он устроен сложным образом. Это относится и к общим закономерностям синтаксической структуры отдельных фраз, и к принципам словосочетания и словорасположения в синтаксических конструкциях. Речь, в частности, идёт о правилах сочетаемости единиц внутри группы терминов (например, согласно одному из таких правил **опустить катет* сказать нельзя, а *опустить перпендикуляр* можно) и о правилах сочетаемости терминов с «обычными» словами русского языка (так, одно из таких правил должно разрешать словосочетание *неравенство треугольника* и запретить **неравенство окружности*). Правила порядка слов в предложениях естественно-подобного языка геометрии должны обеспечить правильность словорасположения любого из терминов относительно всех остальных слов в предложении. Например, порядок расположения идущих друг за другом кванторных групп с разноимёнными кванторами в составе одного предложения менять без изменения исходного смысла нельзя. Так, предложение *Каков бы ни был треугольник, существует вписанная в него окружность* не синонимично предложению *Существует окружность, вписанная в любой треугольник* (одно из предложений истинно, а другое ложно).

2.2. Достоинства и недостатки естественно-подобного языка геометрии

К достоинствам естественно-подобного языка геометрии относится то, что он строится на базе обычного естественного языка, а потому при восприятии геометрических текстов у адресата создаётся психологически приятное ощущение, что они ему понятны. Во многих случаях, однако, такое впечатление иллюзорно, поскольку понимание геометрического текста требует, как минимум, нетривиальных когнитивных операций над ним (см. Крейдлин, Шабат 2011, 2012а) и знания всех входящих в него математических терминов. Ещё одно достоинство текстов на этом языке заключается в том, что они допускают переводы на иностранные естественно-подобные языки, при этом термины переводятся автоматически в силу их интернационального характера.

Из недостатков рассматриваемого языка отметим относительную бедность языковых средств выражения нужных смыслов и ограниченность той предметной области, которую он обслуживает.

Важнейшим его отличием от бытового языка является явная нацеленность на формулировку истин. Между тем, как хорошо известно, на обычном естественном языке люди свободно и охотно выражают не только истинные, но и ложные суждения. Таким образом, естественно-подобный язык — это не простое сужение естественного языка, дополненное некоторыми специальными терминами: из всех нарративных текстов отбираются те, утверждениям в которых можно по определённым правилам приписать истинностные значения. Нацеленность на истинность предполагает точность и однозначность

выражения, поэтому основные тексты геометрии, в которых имеет место синтаксическая или лексическая неоднозначность, подлежат обязательному исправлению, цель которого — уничтожить замеченную неоднозначность.

Ещё одним недостатком рассматриваемого языка является то, что он хуже других языков геометрии поддаётся компьютерной обработке: ведь степень его формализации пока ещё минимальна.

Особую проблему составляет **именование** геометрических объектов и предикатов. Важным свойством рассматриваемого языка геометрии является отсутствие в нём собственных имён и глаголов действий с одушевлёнными субъектами. Различение объектов здесь обычно происходит при помощи местоимений типа *один-другой*, *первый-второй-третий* и т. п. Впрочем, в основные тексты, написанные на естественно-подобном языке, могут входить разные «вкрапления», и этим они нисколько не отличаются от текстов, написанных на обычном русском языке. В русских текстах возможны латинизмы, галлицизмы, германизмы и т. п., слова на иностранных языках, а также единицы невербальных знаковых систем самого разного вида. К таким единицам относятся знаки препинания, разнообразные стрелки, косые чёрточки (так называемые *слэши*), скобки и многие другие знаки.

Заметим, что на сегодняшний день остаётся совершенно не изученной проблема описания употреблений собственных имён в геометрических и, шире, в произвольных математических текстах. Мы бы хотели привлечь внимание читателя к этой важной проблеме.

2.3. Выразимость естественно-подобного языка

Под *выразимостью языка* мы понимаем возможность представить на нём все утверждения о референтной предметной области идиоматичным образом. Далеко не все теоремы геометрии хорошо выразимы на естественно-подобном языке. Например, сформулировать теорему Эйлера $d^2 = R^2 - 2Rr$, связывающую радиусы окружностей, описанной около произвольного треугольника (R) и вписанной в него (r), с расстоянием (d) между их центрами, если и можно, то с большим трудом. Описание класса геометрических утверждений, выразимых на естественно-подобном языке, представляет собой важную и до сих пор открытую междисциплинарную проблему.

3. Язык планиметрических чертежей

3.1. Общие замечания

Особенностью текстов на естественно-подобном языке геометрии является то, что за ними часто стоят понятия, трудно выразимые на этом языке.

Поэтому для их объяснения приходится прибегать к другим языкам геометрии и текстам на них, в частности, к языку планиметрических чертежей. Обратим внимание на то, что слово *язык* здесь используется в переносном смысле, обозначая знаковую систему, причём преимущественно невербального характера. Как и во многих других случаях, когда рассматривается определённая вербальная семиотическая система (естественно-подобный язык геометрии) и функционирующая параллельно с ней невербальная семиотическая система (язык чертежей), возникает проблема их согласования. В подобных случаях А. А. Реформатский (Реформатский 1963) говорил о *конгруировании* таких систем в коммуникативных актах.

3.2. Основные единицы языка планиметрических чертежей

Понятие планиметрического чертежа основано на неопределяемых понятиях *точки* и (евклидовой) *плоскости* — множества точек. Под *геометрическим объектом* в планиметрии понимается подмножество плоскости из некоторого фиксированного класса. Этот класс состоит из элементарных геометрических объектов:

- прямые, лучи, отрезки;
- окружности и дуги окружностей;
- подмножества плоскости, *ограниченные*[2] перечисленными объектами

и сложных, получаемых применением операций пересечения, объединения и дополнения к элементарным объектам.

Важным свойством языка чертежей является отсутствие в нём чисел. Поэтому в нём не выражаются длины отрезков и дуг, площади фигур и величины углов. Однако в нём есть особые средства, позволяющие передать равенства длин и углов (но не площадей!), см. об этом ниже.

Изображения объектов, ограниченных наборами отрезков или дуг, по правилам языка *заштриховываются* (или, в случае использования цвета, *закрашиваются*). Штриховка и цвет являются важными дополнительными семиотическими средствами при построении чертежа. Чертёж может также сопровождаться обозначениями объектов и подписью на специальном языке. Изображение объектов, не содержащихся ни в каком круге[3] — например, внутренней области угла, — требует специальных соглашений. В частности, вводится соглашение о том, что внутренняя область угла заштриховывается или закрашивается.

---

[2] Строгое определение понятия подмножества плоскости, ограниченного набором прямых, лучей, отрезков, окружностей и их дуг, опирается на некоторый достаточно сложный топологический аппарат. Поэтому мы соответствующее определение не приводим.

[3] К сожалению, такие объекты математики тоже называют *неограниченными*. В разделе 3.2 мы употребили выражение *ограниченные прямыми, лучами…*, в котором предикат *ограниченные* имеет две валентности — ‘что ограничено чем’. Другими словами, мы имеем дело здесь с полисемией слов *ограниченный* и *неограниченный*.

3.3. Планиметрические чертежи и их естественно-языковые аналоги

В планиметрии с эллинистических времён сложилось другое понятие чертежа, а именно под чертежом понимались только те изображения геометрических объектов, которые теоретически могут быть построены *циркулем* и *линейкой*. При этом под *линейкой* понималось произвольное средство проведения прямых (линейка без делений), а под циркулем — любое средство проведение окружностей и дуг окружностей.

Бурное развитие компьютерных технологий расширило представление о чертежах, фактически остававшееся неизменным в течение более 2000 лет. Со времён Евклида планиметрические чертежи были *статическими*; сегодня всё чаще встречаются *динамические* чертежи, то есть допускающие *компьютерную анимацию*. Под анимацией, согласно Википедии, понимается « компьютерная имитация движения с помощью изменения (и перерисовки) формы объектов или показа последовательных изображений с фазами движения». Таким образом, динамические чертежи перенесли фокус внимания с вещей на процессы[4]. Кроме того, многие чертежи становятся цветными, и цвет играет в них не только эстетическую, но и смыслоразличительную роль: различие в цветах дифференцирует объекты или их свойства.

В языке чертежей могут использоваться шрифтовые средства (типы, размеры и жирность шрифтов) и специальные знаки, например, разного вида стрелки или маркеры равных отрезков и углов.

В нём имеются аналоги некоторых грамматических классов слов. Так, аналогами существительных являются перечисленные выше имена элементарных геометрических объектов (например, вершины треугольника *ABC*), а вот аналогов прилагательных нет. Это не удивительно, поскольку прилагательное выражает свойство или признак объекта. Зато есть аналоги именных групп прилагательных с существительными. Это определённого вида чертежи, возможно, с некоторыми дополнительными средствами. Например, сочетание *прямой угол* передаётся чертежом:

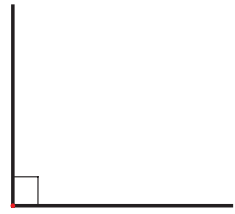


Рис. 1

[4] Подчеркнём, однако, что инновативная культурная геометрическая продукция не препятствует консервации классических статических форм.

Что касается предикатов — глаголов и кратких прилагательных, — то их непосредственных аналогов в языке чертежей нет, однако есть аналоги предикатных групп и более развёрнутых фраз. Приведём примеры некоторых из них.

В языке планиметрии есть несколько синонимичных выражений, передающих одно и то же отношение между точками и прямыми: *Точка лежит на прямой*, *Точка расположена на прямой*, *Прямая проходит через точку*, *Прямая проведена через точку*. В профессиональной математике для передачи этого отношения используется слово *инцидентность*. Говорят, имея в виду одно и то же, *Прямая инцидентна точке*, *Точка инцидентна прямой*, *Точка и прямая инцидентны*.[5] Инцидентность передаётся следующим чертежом:

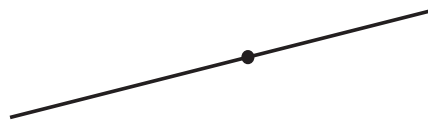


Рис. 2

Ещё одна глагольная группа — это *лежать внутри*.[6] В роли подлежащего у фраз с этой глагольной группой может выступать имя любого ограниченного геометрического объекта, а в роли дополнения — имя любого элементарного геометрического объекта. На языке чертежей отношение *лежать внутри* может выражаться так:

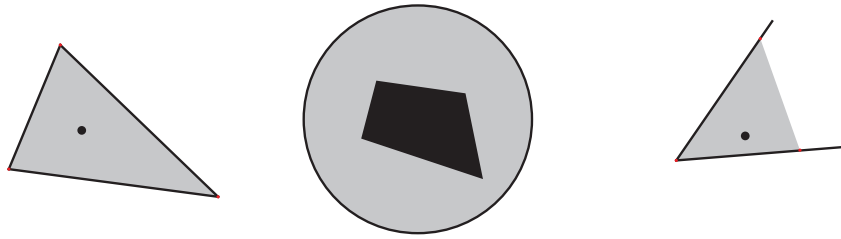


Рис. 3

Приведём пример более сложного выражения на языке чертежей, а именно изображения *середины* отрезка.

Есть два синонимичных способа такого изображения. Первый из них известен со школы:

---

[5] Говорить об инцидентности точки и отрезка, а также точки и луча в математике не принято.

[6] О точке, лежащей на границе геометрического объекта, не говорят, что она *лежит внутри* него.

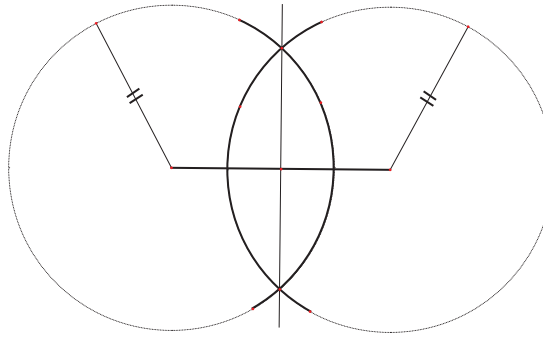


Рис. 4

На этом чертеже с помощью циркуля и линейки построена середина горизонтального отрезка.

Второй способ такой:

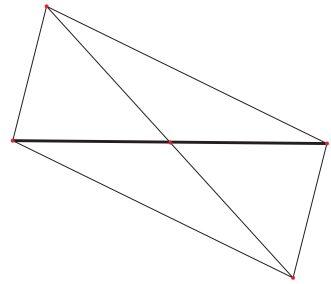


Рис. 5

На этом чертеже отрезок, середина которого строится, является диагональю вспомогательного параллелограмма. Как известно, другая диагональ делит исходный отрезок (изображённый на чертеже жирно) пополам.

При втором способе построения середины отрезка используются только прямолинейные объекты, и в этом мы видим его преимущество перед первым способом. Обратим внимание на то, что в каждом из способов построения участвуют вспомогательные геометрические объекты: в первом случае — это окружности равных радиусов с центрами в концах отрезка, а во втором — параллелограмм с горизонтальным отрезком в качестве одной из его диагоналей.

Проверка *корректности* обоих построений с использованием исключительно языка чертежей невозможна без дополнительных соглашений. Дело в том, что в обоих способах построения неявно участвует квантор общности, который на языке чертежей без таких соглашений не выражается. На языке статической геометрии квантор общности передаётся при помощи построений двух или более *объектов неспециального вида*[7].

---

[7] О понятиях *объект специального вида и объект неспециального* (*общего*) *вида* см. Крейдлин, Шабат 2011.

Корректность построения середины отрезка передаётся чертежами так:

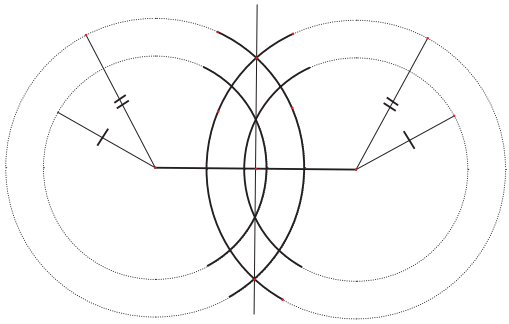


Рис. 6

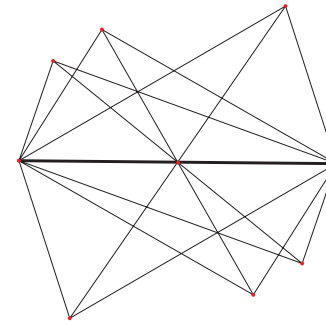


Рис. 7

На этом мы заканчиваем сопоставление двух языков геометрии. За пределами работы остался целый ряд важных проблем, связанных с языком чертежей и его соотношением с естественно-подобным языком, а также с выбором этих языков при коммуникации. Это (а) описание основных лингвистических коррелятов разных уровней, то есть чертежей — аналогов слов, словосочетаний, предложений и текстов; (б) определение некоторых когнитивных операций над чертежами, в частности, выделение значимых фрагментов чертежей; (в) введение особых средств выражения на языке чертежей коммуникативной организации предложений и текстов (среди таких средств — цвет и анимация); (г) разные аспекты, связанные с проблемой переводимости с естественно-подобного языка геометрии на язык чертежей и обратно.

3.4. Выразимость языка чертежей

В данном разделе мы обсудим возможность отобразить на чертеже важнейшие слова и конструкции (синтаксические группы) естественно-подобного языка планиметрии.

Все имена, обозначающие основные геометрические объекты и их части (например, *сторона треугольника*), легко выразимы на языке чертежей, поскольку такие знаки, как *треугольник*, *круг*, *квадрат* и т. п., входят в основной языковой и культурный фонд современного человека. То же можно сказать и о таких синтаксических группах, как *остроугольный треугольник*, *четырёхугольник с взаимно перпендикулярными диагоналями*, *описанная около равнобочной трапеции окружность*. Особенностью выражения всех подобных имён на языке чертежей является то, что они не требуют применения дополнительных средств.

Что касается предикатов естественно-подобного языка планиметрии, то они подразделяются на несколько групп в зависимости от того, требует ли их изображение дополнительных средств, и, если требует, то каких. Примером предиката, не требующего никаких дополнительных средств, является предикат *быть инцидентным*, о котором мы говорили выше. Ещё двумя примерами предикатов той же группы являются *касаться* (в разных значениях):

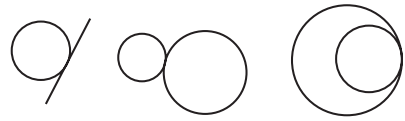


Рис. 8

и *пересекаться* (тоже в разных значениях):

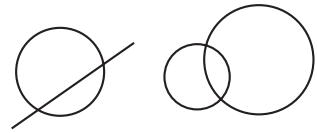


Рис. 9

Изображение предиката *параллелен* в планиметрии опирается на общепринятое соглашение, по которому параллельность определяется «на глаз». Так, чертёж

Рис. 10

выражает параллельность прямых и читается[8] однозначно: *две прямые параллельны*. Вычерчивание большинства остальных предикатов предполагает использование дополнительных средств.

Язык чертежей плохо приспособлен для выражения основных логических связок — конъюнкции, дизъюнкции и отрицания, а о выразимости кванторных конструкций можно сказать следующее. Конструкции с квантором общности с некоторым усилием выразимы на языке статических чертежей. Например, для передачи на этом языке чертеже теоремы *три медианы <u>любого</u> треугольника пересекаются в одной точке* можно построить некоторое достаточно большое количество <u>непересекающихся</u> треугольников, каждый из которых иллюстрирует данную теорему:

Рис. 11

[8] Имеется в виду «читается без использования дополнительных средств — обозначений (которых в рассматриваемых здесь языках вообще нет), цвета, шрифтовой разметки, графических выделений и др.»

Квантор общности прекрасно выразим на языке динамических чертежей, в которых, например, разные треугольники включаются в семейства треугольников, непрерывно перемещающихся по экрану компьютера.

Квантор существования и синтаксические конструкции с ним в языке чертежей выразимы всегда. Дело в том, что геометрия — наука конструктивная. Иными словами, все объекты, существование которых утверждается, допускают построение на плоскости (как правило, при помощи циркуля и линейки в конечное число шагов), которое следует точно сформулированному алгоритму. Как следствие, все построенные геометрические объекты видны на бумаге или на экране компьютера.

Таким образом, одни геометрические тексты на языке чертежей выразимы, а другие — нет. Однако все они выразимы комбинацией языка чертежей с другими языками геометрии, в частности, с естественно-подобным и формальным языками. Уже одно это свидетельствует о важности изучения механизмов взаимодействия разных языков при синтезе и анализе математических текстов, а также в актах коммуникации.

Заключение

Как и в предыдущих наших работах, посвящённых разнообразным проблемам соотношения гуманитарного и естественнонаучного знания, в основание настоящей статьи положены концепты порождения и понимания, математических фактов и текстов.

Выше мы описали два языка геометрии — естественно-подобный и планиметрических чертежей. Поскольку оба языка предназначены для обслуживания одной предметной области, возникают различные вопросы об их соотношении. В частности: *Какой из них предпочтительнее для описания того или иного фрагмента данной области*? *Какой язык лучше понимается и почему*? К ним примыкают вопросы, связанные с переводимостью с одного языка на другой.

Что касается предпочтительности, то мы утверждаем: наилучшее изложение и понимание геометрического текста достигается комбинацией языков. Как и в устной повседневной коммуникации, сочетание вербального и невербального знаковых кодов в письменных математических текстах даёт наилучший эффект. Этому во многом способствует хорошая выразимость геометрических фактов на невербальном языке чертежей.

Основная часть настоящей статьи была посвящена анализу структуры и содержания геометрических текстов. Полное понимание текстов, транслируемых в математической среде, не сводится, однако, к знанию только лексики и грамматики — такое понимание невозможно без учёта прагматики текстов, в частности, особенностей их функционирования и восприятия разными адресатами. Адекватное понимание предметной области всегда достигается рассмотрением её с разных сторон. Важнейшими такими сторонами являются языки, пригодные для её описания. Как удачно выразился Ноа бен

Шиа, «Понимание — это жизнь в доме, где из каждой комнаты открывается свой вид»[9].

Литература

1. *Вежбицкая 1996* — А. Вежбицкая. Язык. Культура. Познание. Москва, 1996.
2. *Гладкий 1997* — А. В. Гладкий 1997. О значении союза ЕСЛИ. Семиотика и информатика, вып. 35. Москва, 1997, 153–183.
3. *Гладкий, Крейдлин 1991* — А. В. Гладкий, Г. Е. Крейдлин. Математика в гуманитарной школе. Математика в школе, 1991. № 6, 6–9.
4. *Звонкин 1990* — А. К. Звонкин. Абстракции с языковой поддержкой. В сб. «Язык и структура знания». Москва, 1990, 86–95.
5. *Корельская, Падучева 1978* — Т. Д. Корельская, Е. В. Падучева. Обратная теорема (алгоритмические и эвристические процессы мышления). Москва, 1978.
6. *Крейдлин 2002* — Г. Е. Крейдлин. Невербальная семиотика: язык тела и естественный язык. Москва, 2002.
7. *Крейдлин, Шабат 2007* — Г. Е. Крейдлин, Г. Б. Шабат. Теорема как вид текста: когнитивные операции и понятность. Вестник РГГУ, 8. Москва, 2007, 102–112.
8. *Крейдлин, Шабат 2011* — Г. Е. Крейдлин, Г. Б. Шабат. Теорема как вид текста II. Когнитивные операции над формулировками теорем. Вестник РГГУ, 11. Москва, 2011, 241–270.
9. *Крейдлин, Шабат 2012а* — Г. Е. Крейдлин, Г. Б. Шабат. Когнитивные операции на пути к пониманию текста. Präsens. Сборник научных трудов. Москва, 2012, 251–265.
10. *Крейдлин, Шмелёв 1989* — Г. Е. Крейдлин, А. Д. Шмелёв. Языковая деятельность и решение задач. Математика в школе, 1989. № 3, 39–45.
11. *Крейдлин, Шмелёв 1994* — Г. Е. Крейдлин, А. Д. Шмелёв. Математика помогает лингвистике. Москва, 1994.
12. *Манин 1977* — Ю. И. Манин. Человек и знак. Природа, 1977, № 5, 150–152.
13. *Манин 2008* — Ю. И. Манин. Математика как метафора. Москва, 2008.
14. *Падучева 1974* — Е. В. Падучева. О семантике синтаксиса. Материалы к трансформационной грамматике русского языка. Москва, 1974.
15. *Пеньковский 2005* — А. Б. Пеньковский. Загадки пушкинского текста и словаря. Опыт филологической герменевтики. Москва, 2005.
16. *Реформатский 1963* — А. А. Реформатский. О перекодировании и трансформации коммуникативных систем. В кн. «Исследования по структурной типологии». Москва, 1963, 208–215.

[9] Ноа бен Шиа «Иаков Пекарь: Простая мудрость для сложного мира». М., 2007: «Рипол-клаассик», с. 141.

17. *Шабат 2008* — Г. Б. Шабат. О симметрии в формулировках теорем. В сб.: «Лингвистика для всех. Летние лингвистические школы 2005 и 2006». Москва, 2008, 203–218.

References

1. *Verzhbitskaja 1996* — Verzhbitskaja A. Jazyk. Kul'tura. Poznanie. Moskva, 1996.
2. *Glalkij 1997* — Glalkij A. V. 1997. O znachenii sojuza ESLI. Semiotika I informatika, vyp. 35. Moskva, 1997, 153–183.
3. *Glalkij, Krejdlin 1991* — Glalkij A. V., Krejdlin G. E. Matematika v gumanitarnoj shkole. Matematika v shkole, 1991. № 6, 6–9.
4. *Zvonkin 1990* — Zvonkin A. K. Abstaktsii s jazykovoj podderzhkoj. V sb. «Jazyk I struktura znanija». Moskva, 1990, 86–95.
5. *Korel'skaja, Paducheva 1978* — Korel'skaja T. D., Paducheva E. V. Obratnaja teorema (algoritmicheskije I evristicheskie protsessy myshlenija). Moskva, 1978.
6. *Krejdlin 2002* — Krejdlin G. E. Neverbal'naja semiotika: jazyk tela I estestvennyj jazyk. Moskva, 2002.
7. *Krejdlin, Shabat 2007* — Krejdlin G. E., Shabat G. B. Teorema kak vid teksta: kognitivnye operatsii I poniatnost'. Vestnik RGGU, 8. Moskva, 2007, 102–112.
8. *Krejdlin, Shabat 2011* — Krejdlin G. E., Shabat G. B. Teorema kak vid teksta II: kognitivnye operatsii nad formulirovkami teorem. Vestnik RGGU, 11. Moskva, 2011, 241–270.
9. *Krejdlin, Shabat 2012a* — Krejdlin G. E., Shabat G. B. Kognitivnye operatsii na puti k ponimaniju teksta. Präsens. Sbornik nauchnyh trudov. Moskva, 2012, 251–265.
10. *Krejdlin, Shmelyov 1989* — Krejdlin G. E., Shmelyov A. D. Jazykovaja dejatel'nost' I reshenie zadach. Matematika v shkole, , 1989. № 3, 39–45.
11. *Krejdlin, Shmelyov 1994* — Krejdlin G. E., Shmelyov A. D. Matematika pomogaet lingvistike. Moskva, 1994.
12. *Manin 1977* — Manin Yu. I. Chelovek I znak. Priroda, 1977, № 5, 150–152.
13. *Manin 2008* — Manin Yu. I. Matematika kak metafora. Moskva, 2008.
14. *Paducheva 1974* — Paducheva E. V. O semantike sintaksisa. Materialy k transformatsionnoj grammatike russkogo jazyka. Moskva, 1974.
15. *Pen'kovskij 2005* — Pen'kovskij A. B. Zagadki pushkinskogo teksta i slovaria. Opyt filologicheskoj germenevtiki. Moskva, 2005.
16. *Reformatskij1963* — Reformatskij A. A. O perekodirovanii i transformatsii kommunikativnyh sistem. V kn. "Issledovanija po strukturnoj tipologii". Moskva, 1963, 208–215.
17. *Shabat 2008* — Shabat G. B. O simmetrii v formulirovkah teorem. V sb.: "Lingvistika dlia vseh. Letnie lingvisticheskie shkoly 2005 i 2006". Moskva, 2008, 203–218.

ПРОСОДИЧЕСКОЕ ЧЛЕНЕНИЕ ЗВУЧАЩЕГО ТЕКСТА: ТЕКСТОВАЯ ЛОКАЛИЗАЦИЯ ДЫХАТЕЛЬНЫХ ПАУЗ[1]

Кривнова О. Ф. (okrivnova@mail.ru)

Московский государственный университет
имени М. В. Ломоносова, Москва, Россия

Ключевые слова: фонетика, устная речь, просодическое членение, интонационно-смысловая дыхательная пауза, текстовая локализация, междикторская вариативность, инструментальный анализ

PROSODIC PHRASING IN SPOKEN TEXT: LOCALIZATION OF BREATHING PAUSES

Krivnova O. F. (okrivnova@mail.ru)

Lomonosov Moscow State University, Moscow, Russia

In this paper we discuss the results of speech breathing research, undertaken to expand an empirical base for modeling of prosodic phrasing in Russian speech, The introductory section provides a brief description of the background, clarifies basic terms, explains the concept of breathing pause (BP) and its correlation with prosodic breaks and prosodic phrasing. In the second section we formulate the problems, discussed in this paper, with the main task to analyze the correlation of BP with the boundaries of the principal text units—paragraphs, sentences, clauses, taking into account the interspeaker variability in reading of the same text. The third section describes the material and methods of experimental analysis with particular attention to the possibilities of the computer detection of BP in a spoken text, as well as to the material adequate to the study. The fourth section outlines the general features of speech breathing in reading of the same text by different speakers. It is shown that one of the most common features is a different number of BPs that speakers make when reading the same text. It was also found that this variability is not related to the gender characteristics of the speakers or their place in the ranking of the best set of readings. Some correlation was found with the individual speech rate — the number of syllables spoken per second. However, despite this variability, all speakers use intonation pauses in the experimental text for breathing rather often. BP part of total intonational pauses averages 62% in the range from 52%

[1] Исследование проведено при поддержке гранта РФФИ 15-06-06103

to 74% by different speakers. The specific use of BP consists in the fact that they reflect the hierarchical structure of the text, with the individual clauses as the basis of it. Namely, text units, the end of which is accompanied by BP, are arranged in the direction of decreasing the probability of BP as follows (in parentheses the frequency of BP averaged by 10 speakers is given): paragraph (100%) > sentence inside a paragraph (94%) > clause inside a sentence (65%) > component in a clause (34%). In conclusion the study is summed up with the implication that BP in prosodic phrasing can serve as a sufficient signal of semantic text boundaries, but interspeaker variability shows that BP is not a necessary indication of them. The differentiating function of this prosodic marker is supported by the fact that BP with different text localization have stable differences in the overall phonetic picture and in such acoustic features as duration and intensity of breathing noise.

Key words: phonetics, spoken language, prosodic phrasing, breathing pause, text localization, interspeaker variability, instrumental analysis

1. Введение

Речепроизводство, как известно, включает три относительно самостоятельных, но тесно взаимодействующих процесса: инициацию (создание воздушного потока и его поддержание в речевом тракте говорящего), фонацию и собственно артикуляцию. Из этих процессов наименее изучен первый, по разным причинам: отчасти из-за преимущественно сегментной направленности речевых исследований во второй половине XX в., отчасти из-за технических трудностей в инструментальном исследовании речевого дыхания и функционирования дыхательной системы в речи.

В то же время исторически потребности дыхания считались одним из главных мотивирующих факторов паузации и интонационно-смыслового членения в устной речи, что вызывало справедливую критику многих лингвистов. В современной фонетике принята точка зрения, согласно которой дыхание пассивно подстраивается под интонационно-смысловые паузы, которые возникают в процессе развертывания высказывания по независимым от потребностей дыхания причинам. Однако эта точка зрения требует, на наш взгляд, уточнения. Возможно, что для говорящего, обладающего автоматизированными интонационно-паузальными стратегиями, потребности дыхания действительно не выступают в качестве мотивирующего фактора паузации: интонационно-смысловые паузы появляются достаточно часто, чтобы в нужный момент сделать вдох. Л. Р. Зиндер пишет в связи с этим: «Человек, у которого органы речи находятся в нормальном состоянии <…>, делает вдох во время пауз между теми или иными синтаксическими единицами, определяющимися смыслом речи. Механизм дыхания предоставляет для этого широкие возможности благодаря постоянному наличию в легких достаточного запаса воздуха, позволяющего при необходимости значительно продлить время фонации» (Зиндер 1979).

Однако локализация вдохов при озвучивании теста не произвольна по отношению к его смысловой и лексико-синтаксической структуре. Судя по имеющимся экспериментальным данным, к сожалению, весьма немногочисленным и отрывочным (Златоустова 1968; Шейкин 1966; Дозорец 1971 а,б; Goldman — Eisler 1968; Grosjean et al. 1979) предпочтительное, хотя и не единственное место речевых вдохов — это конец самостоятельного предложения и конец элементарного предложения — клаузы внутри сложного высказывания. Это говорит о том, что говорящий, по-видимому, располагает особой автоматизированной процедурой речевого дыхания, которая носит рациональный характер по отношению к процессу текстообразования в целом. Возможно, что дыхание как энергетическая база речепроизводства связано с когнитивно-языковыми механизмами через какие-то глубинные психофизиологические структуры. А. Р. Лурия выделяет в качестве одного из основных функциональных блоков мозга единый энергетический блок, который обеспечивает активность коры головного мозга, необходимую для любой психической деятельности. Нарушения в функционировании этого блока проявляются «в одинаковой степени во всех видах деятельности — двигательной, речевой и интеллектуальной, и на всех уровнях коммуникации» (Лурия 1975:58).

Это мнение подтверждается и экспериментами Голдман — Эйслер (1968). На основании наблюдений над речевым поведением людей в норме и под воздействием психотропных лекарств она приходит к выводу, что речевой механизм инициируется к действию и тормозится как целостная система. Речевая интенция, выступая в роли пускового сигнала, активизирует не только когнитивно-языковые процессы, связанные с формированием и вербализацией коммуникативно-смыслового задания, но и дыхательный центр, переводя дыхательную систему в речевой режим деятельности. Степень согласованности разных функциональных процессов в акте речи и конкретные характеристики дыхательной активности зависят от того, насколько говорящий способен контролировать собственное речевое поведение. При оптимальном контроле, предполагающем нормальное для говорящего состояние нервной системы, наблюдается, согласно Голдман — Эйслер, следующее: увеличивается степень соответствия порождаемого текста смысловому заданию, даваемому экспериментатором; возрастает синтаксическая сложность при одновременном увеличении доли грамматических пауз в общем звучании текста; уменьшается частота дыхания при одновременном увеличении речевого объема дыхательного цикла; увеличивается вероятность совпадения вдохов с главными синтаксическими швами, которая достигает 100 % в чтении и 77,6 % в достаточно плавной спонтанной речи.

Согласно идеям, развиваемым в современных моделях порождения речи, тесная корреляция речевых вдохов с границами предложений и клауз кажется не только естественной, когнитивно оправданной, но и необходимой. Однако конкретных эмпирических данных, подтверждающих и иллюстрирующих эту корреляцию, немного (Levelt 1989; Chafe 1994).

2. Задачи исследования

В настоящей работе излагаются результаты исследования, которое было проведено нами для расширения эмпирической базы данных об организации дыхания в устной речи и участии дыхательного фактора в просодической макросегментации звучащего текста.

В исследовании ставились следующие задачи:

- Выявить принципиальные особенности организации речевого дыхания в речи разных говорящих в рамках нейтрального произносительного стиля (при чтении одного и того же текста).
- Оценить степень однородности дыхательного поведения говорящих с целью выделения определенной нормы или предпочитаемых стереотипов, по крайней мере, для режима чтения повествовательного текста.
- Проанализировать согласованность дыхательных пауз с границами основных текстовых единиц — абзацев, самостоятельных предложений, клауз внутри предложений.

Здесь и далее под *дыхательной паузой* (ДП) понимается интонационно-смысловая темпоральная пауза с включенным в нее вдохом.

3. Методика и материал исследования

К сожалению, физиологическая и аэродинамическая сторона речевого дыхания по-прежнему мало доступны для прямого анализа в естественных речевых условиях. В современных исследованиях речепроизводства для получения комплексной картины используются *электромагнитное излучение* и *компьютерная томография*. С помощью этого инструментария можно получить трехмерное изображение речевого тракта и данные об изменении всех его принципиально важных параметров, в том числе дыхательных. Однако, это довольно дорогой инструментарий, и далеко не все исследовательские фонетические центры им располагают. Здесь стоит вспомнить, что еще в 60-е годы XX в. в Институте физиологии им. И. П. Павлова АН СССР была разработана система датчиков, позволяющая регистрировать параллельно работу 11 артикуляторных органов (руководитель работ и изобретатель датчиков проф. В. А. Кожевников). В состав установки входил и плетизмограф, аппарат, с помощью которого можно было регистрировать общую картину речевого дыхания и расхода воздуха при произнесении речевых отрезков. В монографиях (Кожевников и др. 1966; Чистович и др. 1965) приведен ряд интересных результатов, касающихся работы дыхательной системы, которые сих пор сохраняют свою актуальность. К сожалению, установка, разработанная в Институте физиологии, как и многие аналоговые приборы, устарела морально и в настоящее время в научных исследованиях не используется.

Возвращаясь к современности, заметим, что в изучении речевого дыхания не исчерпаны полностью даже самые доступные возможности, которые

предоставляет обычная компьютерная техника, звукозаписывающая аппаратура и программы автоматической обработки речи. Имеющиеся технические средства позволяют, в частности, осуществлять многократное усиление сигнала, в том числе на локальных участках. Если запись речи производится в условиях тихого помещения с использованием высокочувствительного микрофона, можно в большинстве случаев оценить на слух не только наличие вдоха/выдоха в темпоральной интонационной паузе, но и то, через какую полость (носовую/ротовую) осуществляется дыхание. Несколько труднее оценивать на слух глубину вдоха, а она бывает разной, но и такую оценку в определенной степени можно сделать. Современный компьютерный инструментарий, кроме того, делает возможным анализ взаимосвязи между фонетическими параметрами пауз и их акустико-физиологическим заполнением.

Ниже на рис. 1 приведена типовая осциллограмма ДП, полученная из звучащего текста с применением режима локального усиления сигнала. Характерные особенности физического заполнения ДП, хорошо видные на осциллограмме, могут быть положены в основу соответствующей акустической модели и использоваться, например, в распознавании речи для детектирования ДП в звучащем тексте и идентификации текстовых событий, связанных с ними.

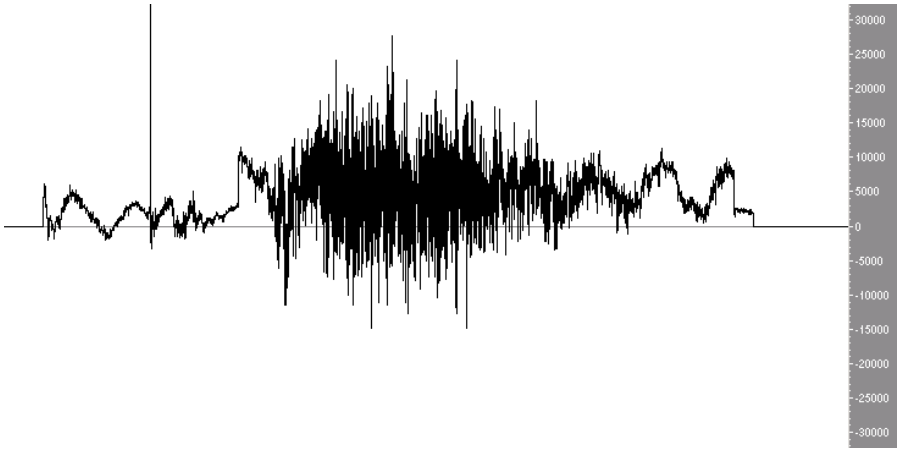


Рис. 1. Осциллограмма усиленной интонационно-смысловой ппаузы с включенным в нее вдохом (типовая акустическая картина)

Материалом исследования служил корпус прочтений связного текста — небольшого современного рассказа о посещении научного учреждения[2]. Текст был прочитан «с листа» десятью дикторами, носителями русского языка с высшим

[2] Текст был взят из методической разработки по составлению текстовых массивов (Штерн 1984), а в качестве основы для него использовался отрывок из книги С. Иванова «Схватка с роботом». М., 1977.

образованием, но без специальной дикторской и лингвистической подготовки. Средняя длительность озвученного текста 3–3,5 минуты. Материал записывался на компьютер (SR 22050 Гц, 16-bit, Mono) в условиях тихой комнаты с использованием высокочувствительного микрофона, что позволило в большинстве случаев без труда определить дыхательный тип пауз в каждом прочтении текста.

Материал был отобран из более крупного массива, включавшего 30 прочтений текста разными дикторами (суммарный объем исходного речевого массива около 400 МБ). При отборе учитывались результаты аудиторского эксперимента по оценке нормативности (приемлемости) разных прочтений текста, который проводился с использованием специально разработанной методики анкетирования аудиторов, подробно описанной в (Кривнова, Чардин 1999). Анкета для опроса аудиторов (их было 6 человек: 4 мужчин и 2 женщины) была составлена таким образом, чтобы отобрать нейтральные, нормативные прочтения. Кроме того, анкета содержала вопросы, специально посвященные оценке правильности паузирования текста (с точки зрения количества пауз, их локализации и длительности, но без акцента на связь с дыханием). Этим оценкам при анализе результатов аудиторской экспертизы был придан больший вес.

Для дальнейшего анализа было выбрано 10 наилучших прочтений, среди которых удачно оказалось 5 мужских и 5 женских — далее они обозначаются соответственно m-i и f-i, где i меняется от 1 до 10 и обозначает место, которое занял диктор в отобранной, лучшей, десятке текстовых прочтений.

Дыхательное заполнение интонационных пауз в прочитанных вариантах текста определялось на слух и визуально по осциллограммам и спектрограммам с использованием звукового анализатора Speech Analyzer — SA SIL, версия 1.5 — 2002.

4. Результаты исследования

4.1. Общие особенности речевого дыхания в прочтениях одного и того же текста разными дикторами

К числу наиболее общих особенностей относится разное количество ДП, которые делают дикторы при чтении одного и того же текста. Специальный анализ, результаты которого приведены в табл. 1, показал, что количество ДП не связано ни с гендерными характеристиками дикторов, ни с их рейтинговым местом в наборе лучших прочтений. Некоторая корреляция обнаруживается с показателями общего темпа произнесения — средней длительностью слога в тексте/количеством слогов, произносимых в сек. Оба показателя определялись по озвученной части текста, без учета внутренних темпоральных пауз; общее количество слогов в анализируемом тексте равно 834. Надо сказать, что по выбранному показателю большинство дикторов, кроме, может быть, f-8, m-9, f-10, читают текст в среднем темпе. По данным (Lenneberg 1967) среднему темпу произнесения соответствуют среднеслоговые длительности 150–170 мс, или

скорость 6,7–5,6 слога в сек. Интересно, что как раз дикторы f-8, m-9, f-10 получили среди лучших чтецов большие штрафы по темпу произнесения и общему количеству пауз. В то же время дикторы со средним темпом речи оказались за пределами типовых показателей средней частоты ДП в речи, которая по литературным данным составляет 16–20 вдохов в минуту (Потапова, Блохина 1986). Для этих дикторов показатели частоты дыхания оказались несколько меньше — 12–15 вдохов в минуту.

Табл. 1. Общие характеристики речевого дыхания и темпа чтения экспериментального текста разными дикторами; данные упорядочены по возрастанию параметра — количество ДП

Дикторы	Штраф за темп	Штраф за кол-во пауз	Сумм. штраф	ДП в % от ТИП	К-во ДП	Длина ДГ в граф. словах		Средн. дл-сть слога в мс	К-во слогов в сек.	Час-та вдохов в мин.
						Ср.	Диапазон			
m-1	0,00	0,29	0,95	62	33	7,6	2–20	167	6,0	12
m-4	0,00	0,57	2,07	57	34	7,8	2–19	165	6,0	13
f-6	0,20	0,00	2,53	62	36	7,6	2–15	147	6,9	14
f-7	0,40	0,00	2,90	52	38	7,0	2–19	173	5,7	13
m-5	**0,20**	**0,29**	**2,15**	**68**	**40**	**6,6**	**2–15**	**153**	**6,5**	**15**
f-10	1,20	0,57	5,10	52	43	6,1	2–12	142	7,0	17
m-9	1,60	1,71	4,48	56	45	5,8	2–13	141	7,1	18
m-3	0,00	0,29	1,79	74	50	5,4	2–12	177	5,7	17
f-2	0,20	0,00	1,03	65	52	5,1	2–9	174	5,8	17
f-8	1,60	0,29	3,89	72	58	4,5	1–12	187	5,4	17

Как видно из табл. 1, несмотря на вариативность в количестве ДП, в целом интонационные паузы используются для вдохов достаточно часто всеми дикторами. Доля ДП в процентах от общего числа темпоральных интонационных пауз (ТИП) составляет в среднем 62% в диапазоне от 52% до 74% по разным дикторам.

Наши данные о связи количества ДП с темпом речи, безусловно, статистически недостаточны и несколько противоречивы; можно лишь высказать предположение, что при медленном темпе количество ДП возрастает и, возможно, свидетельствует об общем замедлении когнитивных процессов и/или более тщательном мониторинге процесса понимания/чтения текста. Этот аспект организации речевого дыхания представляет большой интерес и требует отдельного исследования. На материале английского языка подобное исследование, правда, для пауз в целом, без учета дыхания, было проведено Б. Б. Здороовой (Здоровова 1982). В этом исследовании было обнаружено, что во всех категориях речевого темпа границы самостоятельных предложений и клаузальных компонентов сложных предложений всегда отмечаются физической паузой.

Основные различия наблюдаются в паузации внутри клауз: ускорение темпа приводит к укрупнению единиц макросегментации, а замедление — к их сокращению относительно размеров, типичных для среднего темпа. Эти данные, в целом, совпадают с результатами нашего анализа ДП.

Для общей характеристики речевого дыхания представляет также интерес показатель длины дыхательной группы (ДГ) — цепочки слов, произносимых диктором на одном выдохе. В табл. 1 оценка этого параметра приводится в графических словах, без учета коротких предлогов, союзов и частиц, которые произносились как полные клитики; этот показатель, тем самым, примерно отражает и количество фонетических слов в дыхательной группе.

При увеличении количества ДП в конкретном прочтении средняя длина ДГ, конечно, уменьшается, так как читался один и тот же текст. В связи с показателем длины ДГ нас интересовал не этот бесспорный факт, а абсолютные данные — граничные и средние значения длины ДГ. Из табл. 1 видно, что по показателям верхней границы и средней длины ДГ дикторы делятся на две группы. Для дикторов в нижней части таблицы до m-5 (здесь же примерно проходит и граница по темпу) верхний предел длины ДГ находится в области 12–13 слов, а средняя длина составляет 5–6 слов. Для второй группы эти показатели выше — 15–20 слов и 7–8 слов соответственно. Заметим, что показатели обеих групп не выходят за пределы оптимального объема дыхательного цикла в речи[3]. Источник наблюдаемых различий находится скорее всего в другом: в объеме рабочей памяти дикторов, который действует на организацию речевого дыхания непосредственно или же через корреляцию с размерами интонационно-синтаксических единиц. Этот вывод подтверждается приводимыми ниже данными о дыхательно- текстовых корреляциях (см.табл. 3).

Для удобства дальнейшего изложения введем обозначения для разбиения дикторов на группы, которое наметилось по общим особенностям их дыхательного поведения:

- Группа Г-I — m-1; m-4; f-6; f-7; m-5 — предпочтение крупных ДГ, малое количество ДП.
- Группа Г-II — f-2; m-3; f-8; m-9; — f-10 — избегание крупных ДГ, большое количество ДП.

Далее эти группы условно называются когнитивными[4].

---

[3] Более подробно физиологические особенности оптимальной организации речевого дыхания, наиболее удобной для говорящего, обсуждаются в [Кривнова 2007]. При анализе текстово-дыхательных корреляций, нужно учитывать, что они зависят также от общих фонетических установок говорящего (или читающего): от ориентации на полный/беглый тип произнесения, общую громкость и темп речи.

[4] Интересно, что чтецы, которых аудиторы признали лучшими (первые 4 места в десятке), распределились по когнитивным группам примерно поровну. Возможно, что и аудиторы также делятся на когнитивные группы, и каждый предпочитает близкий себе психостиль чтения.

4.2. Корреляции между локализацией ДП и границами текстовых фрагментов

Экспериментальный текст, несмотря на небольшой объем, содержит разнообразные текстовые единицы — 6 абзацев, 22 самостоятельных предложения и 50 клаузальных единиц, которые являются компонентами самостоятельных предложений. При этом из 22 самостоятельных предложений только 8 являются простыми, моноклаузальными, остальные 14 представляют собой полипредикативные конструкции разного типа. В клаузальной структуре предложений преобладают финитные клаузы — на весь массив из 50 клаузальных единиц, входящих в текст, приходится только 5 нефинитных клауз — причастных и деепричастных. Ниже в табл. 2 в графе 2 без скобок указано количество текстовых единиц данного уровня, которые не являются конечными в единице более высокого ранга, а в скобках — количество единиц, которые завершают единицу более высокого ранга, т. е. отделены нефинальные и финальные составляющие для текстовых фрагментов каждого из интересующих нас уровней с учетом иерархической структуры текста в целом. В графе 3 дается длина текстовых единиц в единицах более низкого уровня: для абзаца в самостоятельных предложениях, а для предложения — в клаузах; приводятся показатели диапазона варьирования и средние значения. В графе 4 длина всех текстовых единиц дается в количестве графических слов; при подсчете этого показателя, как и выше, не учитывались короткие предлоги, союзы и частицы, которые реализуются обычно как абсолютные клитики.

Табл. 2. Композиционно-синтаксическая структура экспериментального текста

Текстовые единицы	**Кол-во в тексте**	**Длина в ед. следующего уровня**	**Длина в граф. словах**	**Конечный знак препинания**
Абзац	6	от 2 до 5	от 19 до 54	.
		Ср. 3,7	Ср. 45,2	
Предложение, самостоятельное внутри абзаца	16 (+6)	от 1 до 5	от 5 до 25	.
		Ср. 2,3	Ср. 12,6	
Клауза внутри предложения	28 (+22)	—	от 2 до 14	, : — ;
			Ср. 5,4	

Даже простое сопоставление таблиц 1 и 2 по параметрам длины ДГ и длины текстовых единиц приводит к заключению, что в дикторских прочтениях можно выделить две схемы организации дыхания при чтении. Одна из них ориентирована на реализацию вдохов на границе между самостоятельными предложениями, другая — на конечных границах отдельных клауз. Это подтверждается и специальным анализом, результаты которого суммированы в табл. 3,

где граница между указанными стратегиями проходит примерно по строке m-5, как и в табл. 1.

Анализ экспериментального материала позволяет выделить особенности дыхательно-текстовых корреляций, общие для всех дикторов. Из табл. 3, приводимой ниже, видно, что все дикторы избегают делать вдохи внутри клаузы: доля ДП здесь у подавляющего большинства дикторов не превышает 20% от общего числа ДП. В то же время темпоральные интонационные паузы (ТИП с вдохом или без него) внутри клаузы представлены достаточно широко. Подсчеты показывают, что они составляют около трети всех темпоральных пауз у каждого из дикторов, и только одна треть из них сопровождается вдохами. Таким образом, незначительная доля ДП внутри клаузы не может быть объяснена «дефицитом» интонационно-смысловых пауз в этом контексте.

Избегая ДП внутри клаузы, все дикторы без исключения делают вдохи после завершения абзацев, независимо от их длины и сложности.

Сходная картина наблюдается и на границах самостоятельных предложений — большинство дикторов делают вдохи в паузах после завершения самостоятельного предложения внутри абзаца. Однако наши данные говорят о том, что здесь для некоторых дикторов (из группы I) становятся существенными такие дополнительные факторы, как сложность и длина предложения, а также предполагаемая степень его связи с последующим продолжением. Это предположение нуждается в дальнейшем исследовании, но едва ли случайно, что в тех немногочисленных случаях, где вдох после предложения отсутствовал, это были простые предложения, начинающие абзац, длиной не более пяти слов. Добавим, что при отсутствии вдоха темпоральная пауза после таких предложений присутствовала во всех прочтениях.

Более сложная ситуация с ДП наблюдается на границах клауз внутри самостоятельного предложения. Здесь количество ДП сокращается, причем в разной степени у разных дикторов. В то же время в этом синтаксическом контексте вероятность темпоральных интонационных пауз (ТИП) также варьирует по дикторам достаточно сильно: в разных прочтениях темпоральными паузами отмечено в сумме от 57 до 96% постклаузальных границ. Данные табл. 3 показывают, что по этому параметру дикторы кластеризуются в две группы с тем же составом, который был намечен выше по параметрам объема дыхательной группы, количества ДП и темпу произнесения (см. табл. 1). Можно полагать, что и источники варьирования наблюдаемой интонационной паузации те же — объем рабочей памяти диктора и скорость протекания когнитивно-языковых процессов. С учетом этого различия в паузальном поведении дикторов использование ими постклаузальных интонационных пауз для вдохов становится однотипным: данные группируются вокруг показателей в 60–70%. Это позволяет предположить, что количество ДП и размеры ДГ определяются прежде всего стратегией интонационного паузирования, характерной для данного диктора, которая, в свою очередь, зависит как от его когнитивных характеристик, так и от клаузальной структуры предложения и длины отдельных клауз.

Табл. 3. ДП на границах текстовых фрагментов разных уровней. Данные дикторов упорядочены по возрастанию числа темпоральных интонационных пауз (ТИП) на границах клауз внутри предложения

Дикторы	ДП на границах клауз (в % от общего числа ДП)	ДП после текстовых единиц разных уровней (в % от общего числа единиц соответствующего типа)					ДП внутри клауз (в % от общего числа ДП)
		После абзацев внутри текста	После самост. предл. внутри абзаца	После клауз внутри предложения	ТИП после клауз внутри предложения в % от общ. к-ва кл.границ	ДП после клауз внутри предложения в % отн. общего к-ва ТИП	
m-1	88	100	82	39	57	69	12
f-7	84	100	82	54	64	67	16
m-4	91	100	100	36	75	48	9
f-6	81	100	91	36	75	48	19
m-5	**83**	**100**	**100**	**43**	**86**	**63**	**17**
f-10	81	100	100	54	86	63	19
m-3	80	100	95	71	86	75	20
m-9	82	100	95	61	89	68	18
f-8	67	100	100	64	89	80	33
f-2	77	100	95	68	96	70	23
Сред.	**81**	**100**	**94**	**53**	**80**	**65**	**19**

Важными параметрами для реализации вдоха на межклаузальных границах внутри предложения являются большая длина произносимой клаузы, ее автосемантичность и ожидание (прогнозирование) развернутого продолжения.

Подводя итоги, можно утверждать, что главным фактором, который влияет на организацию речевого дыхания в репродуцированной речи, является стратегия интонационного паузирования диктора, для которой характерна тенденция к реализации темпоральных интонационных пауз (ТИП) после каждой клаузы в предложении. Однако эта достаточно регулярная тенденция взаимодействует с когнитивными характеристиками дикторов: в результате некоторые дикторы в определенных синтаксических условиях «пропускают» конечные границы произносимых клауз, в то время как другие регулярно реализуют ТИП в определенных точках внутри произносимой клаузы.

Интонационные паузы используются для вдохов достаточно часто — в среднем по дикторам в более чем 60% случаев. Специфика использования таких пауз для речевых вдохов выражается в том, что в организации дыхания находит отражение иерархическая структура текстовых единиц, основание

которой образуют отдельные предикации-клаузы. Текстовые фрагменты, завершение которых сопровождается ТИП с включенным вдохом, упорядочиваются в направлении убывания вероятности вдоха следующим образом (в скобках дается частота реализации вдоха в среднем по 10 дикторам):

Абзац (100%) > самостоятельное предложение внутри абзаца (94%) > клауза внутри предложения (65%) > компонент внутри клаузы (34%).

Когнитивные характеристики дикторов влияют не только на интонационное паузирование, но и на способ организации речевого дыхания в озвученном тексте. В речи дикторов, которые «пропускают» ТИП на границах клауз, встречаются дыхательные группы, совпадающие со сложными предикациями. Для других дикторов типично более частая реализация интонационных пауз и их использование для вдохов, в том числе и внутри клаузы. Эти различия находят непосредственное выражение в таких особенностях дикторского чтения текста, как количество дыхательных пауз, длина и синтаксический состав дыхательных групп.

5. Заключение

Полученные нами результаты подтверждают мнение многих исследователей о центральной роли пропозиции-клаузы в процессах порождения, понимания, озвучивания текста. К сказанному можно добавить, что средняя длина текстовой клаузы (5–6 полнозначных слов), с одной стороны, хорошо согласуется с оценками объема рабочей памяти, а с другой, обеспечивает оптимальный режим работы дыхательной системы у разных дикторов (5–10 слов в дыхательной группе). Можно полагать, продолжая идею А. Р. Лурия о едином энергетическом центре деятельности, что эти корреляции имеют начало в самых глубинных структурах мозга.

К близкому выводу приходили ранее и другие исследователи речевого дыхания на русском материале. Так, в (Дозорец 1971б) указывается, что объяснение объективной необходимости возникновения пауз в речи следует искать, во-первых, в особенностях запоминающей и аналитико-синтетической способности мозга и, во-вторых, в естественном ритме человеческого дыхания.

Из-за наблюдаемой междикторской вариативности реализация ДП в определенных точках звучащего текста не может считаться *необходимым* признаком текстовой границы, однако обнаружение вдоха в интонационной паузе является *достаточным*[5] признаком наличия смысловой текстовой границы, по крайней мере при озвучивании текста в режиме чтения.

Дифференцирующая функция этого типа просодических маркеров в макросегментации текста поддерживается тем, что ДП с разной текстовой локализацией характеризуются устойчивыми различиями в общей фонетической картине и в таких акустических признаках, как длительность и интенсивность

[5] В нашем материале не было ни одного случая реализации вдоха в точке, не оправданной смысловой структурой текста. Возможно, в других дискурсивных режимах такие случаи могут быть обнаружены и должны рассматриваться, видимо, как сбои в правильной организации речевого дыхания.

шума на фазе вдоха (Кривнова 2015). Это создает возможность детектирования ранжированных текстовых границ как в естественном режиме устного дискурса, так и в задачах автоматической обработки звучащей речи, по крайней мере в режиме чтения. Однако этот вопрос требует отдельного обсуждения.

Литература

1. *Дозорец Ж. А.* (1971,а) Проблема членения речи на речевые звенья (синтагмы) и ее разработка в трудах советских языковедов (пятидесятые-шестидесятые годы) // Уч. записки МГПИ им. Ленина. М., N 423.
2. *Дозорец Ж. А.* (1971,б) Эксперимент по определению связи между ритмом дыхания и паузами в речи // Уч.записки МГПИ. Современный русский язык. М.,
3. *Здоровова Б. Б.* (1982) Модификации просодической системы под влиянием изменений темпа речи (экспериментально-фонетическое исследование на материале английского языка). АКД. М.
4. *Зиндер Л. Р.* (1979) Общая фонетика. М.
5. *Златоустова Л. В.* (1968) Некоторые замечания о речевом дыхании // Исследования по речевой информации. М.
6. *Кожевников В. А., Арутюнян Э. А., Бороздин Л. В. и др.* (1966) Методы изучения речевого дыхания // Механизмы речеобразования и восприятия сложных звуков, М.—Л.
7. *Кривнова О. Ф., Чардин И. С.* (1999) Паузирование при автоматическом синтезе речи // Теория и практика речевых исследований (АРСО-99). Материалы конференции. М.
8. *Кривнова О. Ф.* (2007) Фактор речевого дыхания в интонационно-паузальном членении речи // «Лингвистическая полифония». Юбилейный сборник к 70-летию проф. Р. К. Потаповой. М., сс. 424–444.
9. *Кривнова О. Ф.* (1915) Речевое дыхание: локализация и фонетические характеристики дыхательных пауз в репродуцированной речи // Проблемы фонетики, М., Наука, 2015, В.VI, сс. 46–60.
10. *Лурия А. Р.* (1975) Основные проблемы нейролингвистики. М.
11. *Потапова Р. К., Блохина Л. П.* (1986) Средства фонетического членения в немецком и русском языках. М.
12. *Чистович Л. А., Кожевников В. А. и др.* (1965) Речь. Артикуляция и восприятие. М.—Л.
13. *Шейкин Р. Л.* (1966) К механизму возникновения пауз в речи // Механизмы речеобразования и восприятия сложных звуков. М.—Л.
14. *Штерн А. С.* (1984) Артикуляционные таблицы. Методическая разработка для развития навыков аудирования и тестирования слуховой функции. Л.

References

1. *Chafe W.* (1994), Discourse, consciousness, and time. The flow and displacement of conscious experience in speaking and writing. Chicago: University of Chicago Press.
2. *Chistovich L. A., Kozhevnikov V. A. et al.* (1965), Speech. Articulation and perception. [Rech. Artikul'atsija I vosprijatije], M.—L.
3. *Dozorets J. A.* (1971, a). The problem of dividing speech into speech units (syntagms) and its development in the works of Soviet linguists (fifties and sixties years) [Problema chlenenija rechi na rechevyje zvenja i jejo razrabotka v trudah sovetskih jazykovedov], Scientific notes of Lenin Moscow State Pedagogical Institute [Uchonyje zapiski MGPI im. Lenina], N 423, M.
4. *Dozorets J. A.* (1971, b), Experiments to determine the relationship between the rhythm of breathing and pauses in speech [Èksperiment po opredeleniju svjazi mezhdu ritmom dyhanija i pauzami v rechi], Scientific notes of Lenin Moscow State Pedagogical Institute. Modern Russian language [Uchonyje zapiski MGPI im.Lenina. Sovremennyj russkij jazyk], M.
5. *Goldman — Eisler F.* (1968), Psycholinguistics: Experiments in Spontaneous Speech. London and New York: Academic Press.
6. *Grosjean F., Grosjean L., Lane H.* (1979), The patterns of silence: performance structures in sentence production, Cognitive Psychology, V.11, pp. 58–81.
7. *Kozhevnikov V. A., Arutyunyan E. A., Borozdin L. V. et al.* (1966), Methods of speech breathing study [Metody izuchenija rechevogo dyhanija], The mechanisms of speech production and perception of complex sounds [Mehanizmy obrazovanija i vosprijatija slozhnyh zvukov], M.—L.
8. *Krivnova O. F., Chardin I. S.* (1999), Pausing for automatic speech synthesis [Pauzirovanije pri avtomaticheskom sinteze rechi], Theory and practice of speech research (ARSO-99). Proceedings of the conference [Teorija i praktika rechevyh issledovanij (ARSO-99). Materialy konferentsii], M.
9. *Krivnova O. F.* (2007), Breathing factor in prosodic phrasing [Faktor rechevogo dyhanija v intonatsionno-pauzaljnom chlenenii rechi]. «Linguistic polyphony». Anniversary collection for the 70th anniversary of prof. R. K. Potapova [«Lingvisticheskaja polifonija". Jubilejnyj sbornik k 70-letiju prof. R.K. Potapovoj], M., pp. 424–444.
10. *Krivnova O. F.* (1915), Text localization and phonetic characteristics of breathing pauses in reproduced speech [Rechevoje dyhanije: lokalizatsija i foneticheskije harakteristiki dyhateljnyh payz v reprodutsirovannoj rechi], The problems of phonetics [Problemy fonetiki], Science, M., v. VI, pp. 46–60.
11. *Lenneberg E.* (1967) Biological Foundations of Language. N.Y., London.
12. *Levelt W.* (1989), Speaking: from Intention to Articulation. Cambridge: MIT Press.
13. *Luria, A. R.* (1975), The main problems of neurolinguistics [Osnovnyje problemy nejrolingvistiki], M.

14. *Potapova R. K., Blokhina L. P.* (1986), The means of phonetic phrasing in German and Russian languages [Sredstva foneticheskogo chlenenija v nemetskom i russkom jazykah], M.
15. *Sheikin R. L.* (1966), The mechanisms of speech pauses origin [K mehanizmu vozniknovenija pauz v rechi], The mechanisms of speech production and perception of complex sounds [[Mehanizmy obrazovanija i vosprijatija slozhnyh zvukov]], M.—L.
16. *Shtern A. S.* (1984), Articulation tables. Methodical recomendations for developing listening skills and testing of auditory function [Artikuljatsionnyje tablitsy. Metodicheskaja razrabotka dlja razvitija navykov audirovanija I testirovanija sluhovoj funktsii], L.
17. *Zdorovova B. B.*(1982), Prosodic modifications under the influence of changes in the rate of speech (experimental research on a material of English language) [Modifikatsii prosodicheskoj sistemy pod vlijanijem izmenenij tempa rechi (eksperimentaljno-foneticheskoje issledovanije na materiale anglijskogo jazyka)], AKD, M.
18. *Zinder L. R.* (1979), General phonetics [Obshchaja fonetika], M.
19. *Zlatoustova L. V.* (1968), Some remarks on speech breathing [Nekotoryje zamechanija o rechevom dyhanii], Researches on speech information [Issledovanija po rechevoj informatsii], M.

ДИСТРИБУТИВНЫЕ БИМЕСТОИМЕННЫЕ КОНСТРУКЦИИ ТИПА *КТО КУДА*[1]

Кустова Г. И. (galinak03@gmail.com)

Институт русского языка им. В. В. Виноградова РАН;
Московский педагогический государственный университет, Москва, Россия

Местоименные комплексы могут быть единицами системы местоимений (*друг друга*; *что угодно*; *неизвестно где*), а могут быть конструкциями. Конструкция занимает промежуточное положение между свободными сочетаниями и фраземами. В статье рассматриваются биместоименные комплексы (*Разбежались* ***кто куда***), которые имеют дистрибутивное значение. Их первый компонент является дистрибутивом, второй может быть (1) вопросительным местоимением (*Кто* ***когда*** *приехал?* — ‘когда приехал каждый’), (2) неопределенным местоимением (*Занимаются кто* ***чем*** — ‘каждый чем-то своим’), (3) относительным местоимением (*Помогает кому* ***чем*** *может* — ‘каждому тем, чем может’).
В отличие от других конструкций с дистрибутивной семантикой (ср. *Чемоданы попадали на пол*; *Дал каждому по конфете*), дистрибутивные биместоименные комплексы недостаточно исследованы. В работе обсуждаются семантические и синтаксические свойства биместоименных комплексов типа (2) и (3): являются ли местоимения референтными или нереферентными; от чего зависит выбор падежа местоимений.

Ключевые слова: местоимение, референция, дистрибутивная семантика

DISTRIBUTIVE BIPRONOMINAL CONSTRUCTIONS

Kustova G. I. (galinak03@gmail.com)

Vinogradov Russian Language Institute of the Russian Academy of Sciences;
Moscow State Pedagogical University (NSPU), Moscow, Russia

Pronominal complexes may be units of the pronouns’ system (*drug druga*; *chto ugodno*; *neizvestno gde*) or may be constructions. Construction occupies an intermediate position between the free combinations and idioms. This paper discusses the bipronominal complexes (cf.:

[1] Настоящее исследование выполнено при поддержке РНФ, проект № 16-18-02003 «Структура значения и ее отображение в системе лексических и функциональных категорий русского языка».

Razbezhalis' kto kuda), which have distributive meaning. Their first component is a distributive, the second one may be (1) interrogative pronoun (*Kto* ***kogda*** *priehal?* — 'When everyone has arrived'); (2) indefinite pronoun (*Zanimayutsya kto* ***chem*** 'everyone does something'), (3) relative pronoun (*Pomogaet komu* ***chem*** *mozhet* 'He helps everyone with what may').
Unlike other constructions with distributive semantics (cf.: *Chemodany popadali na pol* 'Suitcases fell to the floor'; *Kazhdy poluchil po 10 tsentov* 'They received ten cents each'), distributive bipronominal complexes are insufficiently investigated. The paper discusses the semantic and syntactic properties of bipronominal complexes — (2) and (3) types: are pronouns referential or non-referential; what determines the choice of pronouns' cases.

Key words: pronoun, reference, distributive semantics

Биместоименные комплексы (БМК)

Местоименные комплексы («составные местоимения») традиционно находились на периферии грамматических описаний, что связано с трудностью определения их статуса — они занимают промежуточное положение между словами и конструкциями.

После выхода в свет работы М. Хаспельмата [Haspelmath 1997] возник интерес к неопределенным местоимениям (ср. [Татевосов 2002]; [Падучева 2015]). В число неопределенных местоимений включают и комплексы, состоящие из К-местоимения и дополнительного элемента — так называемые квазирелятивы, или свободные релятивы, типа *кто угодно / куда попало / что положено / как следует* — и так называемые амальгамы типа *неизвестно где / черт знает кто / абы как* (ср. [Тестелец, Былинина 2005]). В русской грамматической традиции такие образования рассматривались, в первую очередь, с точки зрения синтаксического статуса (считать ли их редуцированными предложениями или членами предложения, ср. [Егорова 1967]) и семантики (см. [Откупщикова 1984], [Шелякин 1986]). В современной лингвистической литературе свободные релятивы и амальгамы описываются именно как местоимения, ср. [Тестелец, Былинина 2005], [Евтюхин 2008], [Кустова 2015].

К относительно хорошо изученным можно отнести также собственно местоименный комплекс — взаимно-возвратное местоимение *друг друга* ([Падучева 1985], [Шелякин 1986: 32], [Тестелец 2001: 46–47], [Гайнутдинова 2012]). Однако *друг друга* — уникальная единица местоименной системы (хотя у нее есть аналоги типа *один другого*). Между тем в системе местоимений существует целая группа биместоименных конструкций, свойства которых практически не описаны, хотя, безусловно, заслуживают внимания, — это конструкции типа *кто куда*, *кто с кем*, *когда как* и под. Мы будем называть их биместоименными комплексами (сокращенно — БМК), ср.: [*С работы уходят*] [***кто когда*** *захочет*]. Сразу оговоримся, что количество местоимений в комплексе может быть и больше двух (*Бегают кто когда куда захочет*; *Выясни, кто с кем о чем разговаривал*), но такие конструкции встречаются редко, и их устройство аналогично БМК, поэтому дальше мы будем рассматривать только БМК.

БМК образованы так называемыми К-местоимениями. К К-местоимениям принято относить местоимения *кто*, *что*, *где*, *когда* и т. д. ([Откупщикова 1984], [Апресян, Иомдин 2010)]), которые могут функционировать как вопросительные (***Кто*** *пришел?*); относительные (*Тот*, ***кто*** *пришел*), неопределенные (*Если* ***кто*** *придет*, *скажи* = ‘кто-нибудь’). В образованных К-местоимениями комплексах реализуется еще одно значение, которое и будет нас интересовать, — ДИСТРИБУТИВНОЕ, ср. *Разбежались* ***кто куда*** = ‘каждый в свою сторону’; *Узнайте*, ***кто когда*** *приехал* = ‘когда приехал каждый [из известного множества]’. Дистрибутивное значение не выделяется в классификациях, не фигурирует в грамматических описаниях и словарях — возможно, потому, что оно не существует в изолированном употреблении, а существует только в составе биместоименных конструкций, а они не считаются целостными единицами типа *друг друга*.

В литературе конструкции с двумя местоимениями описываются с разных точек зрения.

Авторов, которые работают в парадигме генеративной грамматики или формальной семантики, интересуют такие проблемы, как линеаризация местоименных элементов в вопросах ([Bošković 2001; 2002]), либо способы формального представления конструкций с двумя местоимениями — также в вопросах, ср.: *Who ate what?*; *Which book did which student read?* (см. [Kotek 2007], [Dayal 2005]).

Из отечественных исследований можно отметить работу [Шведова 1998], в которой конструкции *кто где*, *кто куда*, *кто про что* и под. упоминаются как элементы системы русских местоимений (они названы «устойчивыми речениями со значением неопределенной и рассредоточенной множественности»), однако, во-первых, рассматриваются только пары с компонентом *кто*, — поскольку эти конструкции, наряду с другими, обсуждаются в рамках явления «человекоцентризма»; во-вторых, речь идет только об одном случае употребления БМК — о функционировании в ответной реплике:

> *— Легко справились с болезнью? — Кто как; Большие деньги получили? — Кто сколько; — Чем же отблагодарили? — Кто чем* (примеры из [Шведова 1998]).

Между тем БМК образуются не только на базе *кто* (ср. *когда как*; *куда сколько*; *что откуда*) и употребляются не только в ответных репликах, но и в целом ряде других конструкций.

Ниже мы рассмотрим некоторые свойства дистрибутивных БМК.

Дистрибутивные конструкции

Дистрибутивное значение и конструкции с дистрибутивной семантикой широко представлены в языке. К ним относятся предложения с глаголами дистрибутивного способа действия, ср. *Чемоданы попадали с полки*; *Все сотрудники переболели гриппом*; дистрибутивная конструкция с предлогом *по*, ср.: *Дети получили по яблоку* (см. [Лингвистика конструкций 2010: 184–218]).

Что касается системы местоимений, то здесь до сих пор выделялись только лексические дистрибутивы — например, кванторное местоимение *каждый* (в отличие от *любой*), которое свободно сочетается с дистрибутивной конструкцией: *Каждый получил по яблоку*.

Примечание. В [Откупщикова 1984] дистрибутивными считаются местоимения *каждый* и *всякий*, однако *всякий*, во-первых, стилистически окрашено, а во-вторых, чаще употребляется в адъективных значениях типа ‘разный’ (*Показывал всякие фокусы*; *Накупили всяких игрушек*) или в значении низкой оценки (*Будут еще всякие менеджеры мне указывать, я сам начальник*).

Кроме того, в работах по местоимениям упоминается дистрибутивная референция, см. [Шмелев 2002: 98–100], и дистрибутивная интерпретация именных групп [Падучева 2015].

В данной работе мы хотим привлечь внимание к языковым особенностям группы грамматикализованных дистрибутивных конструкций с двумя местоимениями.

Дистрибутивные БМК

Подобно квазирелятивам (*кто / что / где… угодно*; *кто / что / где… попало*; *кто/ что / когда… хочешь*) и амальгамам (*неизвестно кто / что / где / зачем…*; *не пойми кто / что / какой…*), дистрибутивные комплексы образуют серии (ниже приводятся не полные списки, а лишь иллюстрации): *кто кого* (*кто кому*, *кто с кем…*; *кого с кем*, *кому с кем*, *о ком с кем*; *кого о ком*, *кому о ком*; *с кем о ком…*); *кто что* (*кто чем*, *кто о чем* и т. д.); *кто где*; *кто сколько* (*кому сколько*); *когда кто*, *когда где*, *когда как* и др.

Первое местоимение в дистрибутивных БМК всегда имеет дистрибутивную семантику, поэтому далее будем называть его дистрибутив (речь идет об иерархически «первом» — т. е. главном, — элементе; линейно дистрибутив может стоять и на втором месте — особенно если он неодушевленный [*что*] или адвербиальный [*где*, *когда* и под.], см. ниже).

Второе местоимение комплекса может иметь разный статус — в зависимости от того, в каком типе предложений встречается БМК; оно может быть вопросительным, неопределенным или относительным:

(1) *Вопросительное местоимение — в вопросительных конструкциях.*

Дистрибутивные комплексы встречаются в собственно вопросительных предложениях: *Кто когда приехал?* (‘когда приехал каждый’) — и в косвенном вопросе (в придаточном): *Выясни, кто когда приехал*.

Первое местоимение, т. е. дистрибутив (в наших примерах — *кто*), не входит в сферу действия оператора вопроса, второе является вопросительным. Т.е.

предложение *Выясни, кто когда приехал* не содержит смысла ‘выясни, кто приехал’, а означает ‘выясни, когда приехал каждый из известных / ранее установленных’.

Два вопросительных местоимения могут обозначать и два отдельных вопроса. В идеале это должна быть сочинительная конструкция: *Выясни, кто и когда приехал*, но в разговорной речи сочинительный союз может отсутствовать. Однако семантические и референциальные признаки вопроса при этом все равно сохраняются: в собственно вопросительном предложении не известно ни первое множество (кто приехал), ни второе множество (моменты времени), и говорящего интересуют оба множества. В дистрибутивной конструкции первое множество известно, и требуется установить второе множество и сопоставить его элементы элементам из первого множества: — *Вот список постояльцев. Выясни, кто когда приехал*.

Вопросительные конструкции мы далее рассматривать не будем.

(2) *Неопределенное местоимение — в изоляте.*

Дистрибутивные комплексы могут употребляться в ответной реплике (в ответе на вопрос):

— *Где вы отдыхали?*
— *Кто где* (‘каждый где-то отдыхал; каждый в своем месте’)*;*
— *Куда вас направили?*
— *Кого куда* (‘каждого куда-то направили’)*;*
— *Кто с ребенком сидит?*
— *Когда кто* (‘каждый раз кто-то / кто-нибудь сидит’)*;*
— *Сколько они получают?*
— *Кто сколько* (‘каждый какую-то сумму’).

Дистрибутивные комплексы с неопределенным вторым местоимением мы называем изолятами, поскольку они могут употребляться абсолютивно, без глагола и других элементов конструкции, — в отличие от квазирелятива (*Зрители садятся кто где* ***хочет***). Соответственно, присоединение «правого» элемента (предиката) превращает изолят в квазирелятив, ср.: *Получают кто сколько* vs. *Получают кто сколько* ***попросит***.

Изоляты могут употребляться и в составе предложения — будем называть их включенными изолятами (этот термин, конечно, содержит противоречие, но мы хотим подчеркнуть, что к изолятам неприсоединим правый контекст; левый контекст они допускают): *Десятки бойцов* ***сорвались кто откуда*** *и, наматывая солдатские ремни на руку, тотчас ринулись к мосту* [Юрий Буйда. Красавица Му (1998)].

Изолят может иметь расширение — конкретизацию объектов, обозначенных неопределенным местоимением: *Дети получили* ***кто что****:* ***кто*** *машинку,* ***кто*** *куклу*; — *Никому тут моя наука не нужна. Институт давно закрыли, сотрудники занимаются* ***кто чем***. ***Кто*** *уехал,* ***кто*** *на рынке торгует…* [Евгений Чижов. Перевод с подстрочника (2012)].

Неопределенное местоимение в изоляте обычно референтное. Нереферентной интерпретации местоимения, а значит, и контекстам снятой утвердительности (о контекстах снятой утвердительности см. [Падучева 2013]), изоляты сопротивляются:

?*Прячьтесь кто куда;*
?*Берите кто что и спускайтесь к машине;*
?*Если они получат кто сколько, смогут сделать взнос.*

(3) *Относительное местоимение — в квазирелятиве.*

Квазирелятивы (свободные релятивы) — безвершинные относительные придаточные (см. [Тестелец, Былинина 2005]), которые возникли в результате редукции местоименно-соотносительного придаточного (см. [Кустова 2015]): *Отдыхали там, где собирались → Отдыхали где собирались*; *Сообщи тому, кому положено* [*сообщать*] *→ Сообщи кому положено*.

Конструкции типа *Берите кому что нравится* образованы на базе квазирелятива (ср. *Берите что нравится / что хотите / что предложат*) путем включения второго местоимения в функции дистрибутива: ‘берите то, что каждый хочет’ → *Берите кто что хочет*. Будем называть их дистрибутивными квазирелятивами или, для краткости, просто квазирелятивами. Они включают три компонента: два материально выраженных — дистрибутив (в нашем примере — *кто*) и релят (относительное местоимение; в нашем примере — *что*) — и один подразумеваемый — это опущенный коррелят (указательное местоимение): *берите* [**то**] *что хотите*.

Релят может быть как нереферентным: *Хватайте кто что успеет* (‘что-нибудь’), так и референтным: *Похватали кто что успел* (‘каждый — что-то конкретное’). Большинство примеров дистрибутивных БМК в НКРЯ — с референтным релятом.

Дистрибутивные БМК: свободные сочетания, единицы или конструкции?

Хотя, на первый взгляд, дистрибутивные БМК образуются «регулярно», т. е. теоретически можно соединить любые два местоимения, в действительности они подчиняются целому ряду ограничений. Дистрибутивные БМК занимают промежуточное положение между свободными сочетаниями типа сочинительных конструкций *кто и сколько*, *куда и зачем* и под. и грамматикализованными сочетаниями типа *что попало* или *друг друга*, которые функционируют как отдельные единицы местоименной системы. В силу этого дистрибутивные БМК нельзя рассматривать как свободные сочетания, а следует считать конструкциями (в смысле [Лингвистика конструкций 2010]).

Конструкция имеет устойчивый характер (обычно не допускается замена компонентов), идиоматичную семантику, грамматические ограничения. Все это относится и к дистрибутивным БМК.

(а) Многие БМК имеют статус фразем или употребляются в составе фразем: *Когда как*; *Кому какое дело*; *Кто во что горазд*; *Кто на что учился*; *Откуда что взялось*; *Еще посмотрим, кто кого*.

(б) БМК не допускают замены элементов на другие типы местоимений:

*Дети получили в подарок кто что — *Дети получили каждый что; *Дети получили каждый что-то; *Дети получили кто что-то.*

Кажется, что квазирелятив допускает замену дистрибутива на *каждый*, но, во-первых, при этом необходимо изменить порядок элементов: *Дети получили в подарок кто что хотел — ?Дети получили каждый что хотел → Дети получили что каждый хотел*; во-вторых, конструкция *Дети получили что каждый хотел* звучит неестественно и в Корпусе не встречается.

(в) Выбор падежной формы дистрибутивных БМК подчиняется определенным правилам, как и выбор форм в сочетаниях *друг друга* или *сам себя* (см. ниже).

(г) Наконец, элементы дистрибутивных БМК находятся в определенной иерархии (см. ниже).

Далее мы рассмотрим изоляты и квазирелятивы. Несмотря на внешнее сходство местоименных пар, у них разная семантика и разные синтаксические свойства.

Изоляты

Сначала — несколько замечаний по семантике дистрибутивных БМК модели *Разбежались кто куда*.

Обычно дистрибутивные конструкции обобщают одинаковые или похожие ситуации: *Каждый получил по 100 рублей*; *Все чемоданы попадали на пол*; *Все сотрудники переболели гриппом*. Даже если в ситуациях есть различия, ср. *У нас переболела половина сотрудников* (возможно, разными болезнями), — акцентируется сходство, в данном случае — сам факт болезни.

Дистрибутивный изолят, напротив, акцентирует различия: *Сколько они получают? — Кто сколько* значит ‘все по-разному’, ср. возможное расширение: *кто сколько: кто 100 рублей, кто 200* — и невозможное: **кто сколько — по 100 рублей*; ср. также: *Разъехались кто куда* — ‘в разные места, каждый в свое место’, *Вспоминали кто о чем* — ‘каждый о своем’. Таким образом, у разных дистрибутивных конструкций разные функции: одни акцентируют сходство ситуаций, другие — различие.

Вторая семантическая особенность изолятов: о референте неопределенного местоимения не сообщается ничего конкретного, изолят просто констатирует различия: *Дети получили в подарок кто что, а взрослые — телефоны* — в этом предложении нет конкретной информации о подарках детям. При этом в контексте дистрибутива неопределенное местоимение имеет иную интерпретацию, чем обычное неопределенное местоимение типа *что-то, куда-то* и под., ср.:

Друзья уехали куда-то — говорящий НЕ знает, куда именно уехали друзья;

Друзья разъехались кто куда — уехали в разные места, причем неопределенными (неизвестными) эти места являются только для адресата, что же касается говорящего, то он ЗНАЕТ, куда именно уехал каждый, и может это сообщить: *Друзья разъехались кто куда: Таня в Новосибирске*, *Наташа в Томске*, *Миша в Петербурге*.

С синтаксической точки зрения изолят находится в той же клаузе, что и предикат, от которого он зависит, и падежи его местоименных элементов определяются управлением этого предиката: *Дети получили кто что / Детям подарили кому что*. Неопределенное местоимение (далее сокращенно — НМ) *что* соответствует дополнению глагола, а дистрибутив является чем-то вроде плавающего определителя и дублирует падеж кореферентной ИГ (именной группы), ср. конструкции с плавающими определителями типа *сам*: ***Дети*** *наливают суп* ***сами*** */* ***Детям*** *придется наливать суп* ***самим***.

При этом есть два важных отличия от плавающего определителя: (1) плавающий определитель может быть расположен как контактно (***Дети сами*** *наливают суп*), так и дистантно (***Дети*** *наливают суп* ***сами***); элементы БМК всегда расположены контактно (*Награбленное рассовали* ***что куда***); (2) плавающий определитель относится к существительному, дистрибутив может быть и адвербиалом (*Семинары проводим* ***когда*** *где*).

Существует ли иерархия дистрибутива и НМ, т. е. какое из двух местоимений интерпретируется как дистрибутив, а какое — как неопределенное, и должен ли дистрибутив иметь более высокий ранг, чем НМ?

На основании приведенных примеров можно сделать вывод, что дистрибутив действительно имеет:

- более высокий синтаксический ранг, например, дистрибутив — подлежащее, а НМ — дополнение: *Дети получили кто* [Им.] *что* [Вин.]
- или более высокий семантический ранг, например, у дистрибутива — одушевленный референт, у НМ — неодушевленный: *Детям подарили кому что*.

Интересно, что даже если НМ — подлежащее, дистрибутивом все равно является местоимение с одушевленным референтом: *Детям досталось кому* [одуш. Дат.] *что* [неодуш. Им.] — *кому кукла*, *кому машинка*. Вообще, кажется существенным, чтобы в дистрибутивной конструкции главным было лицо, ср.:

— Где работают ваши выпускники?
— Кто где: кто в школе, кто в фирме, кто на радио;
— Где работают ваши выпускники?
*— *Где кто: где Витя, где Маша.*

(не случайно Н. Ю. Шведова рассматривает конструкции *кто где*, *кто куда* в разделе «человекоцентризм» [Шведова 1998]).

Однако, на самом деле, одушевленное местоимение не всегда является дистрибутивом. В дистрибутивном БМК есть иерархия, но она определяется коммуникативной структурой текста. Дистрибутивом становится известное, данное, тема, а НМ имеет статус ремы:

— Кто дежурит на объектах? — Где кто: где прораб, где главный инженер (‘на каждом объекте — свой дежурный’) — здесь данным являются места, а новым — люди, и дистрибутивом становится обстоятельство места (другое дело, что известным, данным часто является подлежащее или иной личный субъект).

Ср. также БМК с двумя неодушевленными компонентами:

— Где продаются фрукты? — Что где: яблоки на рынке, ананасы в магазине — при обратном порядке элементов БМК их интерпретация сохраняется: *— Где продаются фрукты? — Где* [НМ] *что* [дистрибутив]*: яблоки на рынке, ананасы в магазине;*

— Что продается на рынках? — Где что / что где: где рыба, где фрукты.

Дистрибутивные квазирелятивы

У квазирелятива есть и семантические, и синтаксические отличия от изолята.

У квазирелятива нет требования, чтобы все элементы множества различались. Семантика квазирелятива основана на отождествлении: *Дети получили кто что хотел* = ‘[то] что каждый хотел’ — например, если три мальчика из пяти хотели машинку, то среди подарков будет три машинки.

Если в изоляте объекты, к которым отсылает неопределенное местоимение, известны говорящему (*Друзья разъехались кто куда* — говорящий знает, куда именно), то в дистрибутивном квазирелятиве референт релята может быть и неизвестен говорящему, но в какой-то степени охарактеризован, т. к. задан предикатом зависимой клаузы: *А уж когда разделим — меняйтесь кто с кем* ***хочет*** [Сергей Иванов. Марш авиаторов // «Звезда», 2002]; *Подайте кому сколько* ***не жалко***; *Можете использовать оборудование кому когда* ***потребуется***; *Награбленное рассовали что куда* ***влезло*** *— и бежать* и т. д.

В квазирелятиве две клаузы, и формы падежей обоих местоимений определяются не предикатом главной клаузы, а предикатом зависимой: *Дети получили* ***кто что хотел*** — падежи *кто* и *что* определяются в рамках конструкции ‘кто хотел получить что’, ср.: *Дети получили* ***кому что нравилось***.

В приведенных примерах дистрибутив имеет более высокий ранг, чем релят, — это либо подлежащее: *Дети получили* ***кто*** *что хотел* (что **каждый** хотел), либо одушевленный референт: *Дети получили* ***кому*** *что нравилось* (что **каждому** нравилось). Как и в случае с изолятом, возникает вопрос: является ли это правилом или частным случаем? Такая тенденция действительно прослеживается, и связана она опять-таки с коммуникативной структурой: у иерархически более высоких участников больше вероятность оказаться в теме. Однако в общем случае дистрибутивом становится известное, названное, введенное в рассмотрение, поэтому дистрибутивом может быть и адвербиальный элемент БМК, имеющий статус данного:

— *С кем вы оставляете ребенка* [‘каждый раз’ — множество моментов времени имплицировано узуальным значением вида]*?* — ***Когда*** *с кем получится* (**каждый раз** с тем, с кем получится’)*;* — *Кто у вас дежурит на объектах?* — ***Где*** *кто может* (‘**на каждом** тот, кто может’).

Что касается правил выбора падежей, то они более сложные, чем в изоляте, т. к. дистрибутивный квазирелятив, как уже говорилось, находится в отдельной клаузе.

При этом падеж каждого элемента БМК подчиняется своим собственным правилам.

Падеж релята зависит от опущенного (реконструируемого) коррелята. Об этом мы подробно говорили в другой работе (см. [Кустова 2015]), сейчас кратко сформулируем основной принцип: чтобы можно было опустить коррелят (т. е. получить свободный релятив), падежи коррелята и релята должны совпадать: *Кто выучил, получит «пять»* [Им. (*тот*) = Им. (*кто*)]; ?*Кто выучил, поставят «пять»* [Дат. (*тому*) ≠ Им. (*кто*)]. Чтобы падежи актантов совпадали, в каждой клаузе должны быть либо предикаты с одинаковым управлением, либо один и тот же предикат (в том числе материально не выраженный) — тогда управление заведомо будет одинаковым. Именно поэтому так много квазирелятивов включают модальный компонент, который присоединяет инфинитив, совпадающий с глаголом главной клаузы: ***бери*** [то] *что хочешь* [**взять**]; ***предупреди*** [того] *кого сможешь* [**предупредить**] и т. п.

Аналогичное правило действует по отношению к реляту и в дистрибутивных конструкциях:

(а) либо совпадает управление разных предикатов: *Съели кто что принес* (‘съели то — принес что’); *Дети сидят за партой кто с кем дружит* (сидят с тем — дружат с кем);

(б) либо в зависимой клаузе есть повторяющийся (невыраженный) предикат: *Дети сидят за партой кто с кем хочет* [сидят с теми — (хотят) сидеть с кем]; *Соседям-погорельцам будем помогать кто чем может* [помогать тем — (может) помочь чем].

Примечание. Из этого правила есть по крайней мере два исключения:

- падеж может не совпадать, если совпадают формы: *Остальные получили кому что досталось / Дети получили кому что понравилось* [получили то (Вин.), что (Им.) досталось / понравилось];
- предложный падеж не требует совпадения: *Дети получили* [**то**] *кто* ***о чем*** *мечтал*.

Есть также различные ограничения (ср.: ?*Дети любят кто что делает /* ?*Детям нравится кто что делает* — в значении ‘нравится то, что каждый делает’; ?*Детям нравится кто что получил в подарок / кому что подарили*), на которых мы сейчас не можем останавливаться.

Падеж дистрибутива подчиняется другим правилам.

Если референт дистрибутива в главной клаузе — подлежащее, то дистрибутив может стоять в разных падежах — т. е. дистрибутив стоит в том падеже, какого требует предикат зависимой клаузы (квазирелятива), и не должен совпадать с падежом кореферентной ИГ в главной клаузе, т. е. с именительным падежом подлежащего: *Дети получили* ***кто*** *что* ***хотел***; *Дети получили* ***кому*** *что* ***понравилось***; *Дети сделали* ***кого*** *о чем* ***просили***; *Мы дежурим* ***кого*** *с кем* ***поставят***.

Если кореферентная дистрибутиву ИГ в главной клаузе находится в подчиненной позиции (дополнение), то дистрибутив в зависимой клаузе должен стоять в том же падеже, что и ИГ в главной (при этом не обязательно, чтобы в обеих частях был один и тот же предикат, достаточно, чтобы у разных предикатов было одинаковое управление, — но может быть и повторяющийся предикат), ср.:

> ?***Детям*** (Дат.) *подарили* ***кто*** (Им.) *что хотел — **Детям*** (Дат.) *подарили* ***кому*** (Дат.) *что понравилось;*
>
> ?*Он всегда помогает* ***кто*** *о чем попросит* [помогает **тому** — просит **кто**] — *Он всегда помогает* ***кому*** *чем может* [помогает **каждому** — **кому** может помочь]*;*
>
> ?***Нас*** *поселили* ***кто*** *с кем хотел — **Нас** поселили* ***кого*** *с кем смогли;*
>
> ?***Их*** *арестовали* ***кто*** *где был — **Их** арестовали* ***кого*** *где застали.*

(первое предложение из приведенных пар не является абсолютно неправильным, но имеет просторечный характер, ср. примеры из [Егорова 1967: 89]: — *Что с рыбой-то делать? — Отдай у кого кошки есть*; *Надо прислушиваться что люди советуют*; *Найди мне кто вяжет*).

Заключение

Итак, наряду с хорошо изученными дистрибутивными конструкциями типа *дал каждому по конфете* существует большой класс дистрибутивных БМК, которые имеют свою структуру и свои законы функционирования. Анализ материала показал, что дистрибутивные биместоименные конструкции имеют как сходства, так и различия.

Общие признаки:

1. Дистрибутивные БМК являются результатом работы единого языкового механизма, вводящего в исходную конструкцию с местоимением ещё один местоименный элемент с дистрибутивным значением (*узнай, когда они приехали → узнай, кто когда приехал*; *уехали куда-то → уехали кто куда*; *гуляют где хотят → гуляют кто где хочет*). Этот элемент в БМК является иерархически главным и обычно (но не всегда) линейно первым.

Дистрибутив позволяет представить некоторое множество (множественный актант) не как совокупное, а как расчлененное, дифференцированное, ср.: *Детям что-то подарили* — не исключается, что это были одинаковые подарки или даже один подарок на всех vs. *Детям подарили кому что* — ‘каждый получил отдельный подарок; подарки разные’.

2. Главное местоимение в БМК, которое по своей исходной семантике не являлось дистрибутивным (как лексический дистрибутив *каждый*), а являлось вопросительно-относительным, приобретает дистрибутивную семантику. Семантика и функции другого местоимения определяются тем, в какую объемлющую конструкцию оно входит: в вопрос (вопросительное), в изолят (неопределенное) или в квазирелятив (относительное).
3. Элементы дистрибутивных БМК всегда расположены контактно: *Узнай, кого куда поселили* — *?Узнай, кого поселили куда*; *Сотрудники получают за работу кто сколько* — **Сотрудники кто получают сколько* — **Сотрудники сколько получают кто*; *Дети сидят за партой кто с кем дружит* — **Дети сидят с кем дружит кто*.

   Несмотря на контактное расположение, элементы БМК не образуют единой группы, что видно из экспликации: *Выпускники журфака работают кто где: кто в газете, кто на радио*; *Арестовали кого где застали* — ‘каждого арестовали там, где его застали’.
4. В дистрибутивных БМК существует иерархия, которая имеет коммуникативную природу: дистрибутивом является данное (обычно это лицо, но не всегда), второй местоименный элемент имеет статус нового.

С другой стороны, поскольку БМК встроены в разные типы предложений, между ними существуют различия.

Изоляты связаны с одной клаузой, и их формы заданы элементами этой клаузы. Если местоимения в БМК субстантивные, ср. *Детям подарили кому что*, падеж неопределенного местоимения определяется предикатом (*подарили что*), падеж дистрибутива (по аналогии с плавающим определителем) дублирует падеж именной группы (*детям … кому*), который, в свою очередь, также зависит от предиката. Адвербиальные компоненты БМК, ср. *Семинары проводим когда где*, тоже связаны с предикатом.

Квазирелятивы связаны с двумя клаузами, что естественно, т. к. квазирелятив — безвершинное относительное придаточное. Падежи местоимений в БМК зависят как от элементов подчиненной клаузы, в которой находится БМК, так и от элементов главной клаузы.

Литература

1. *Апресян Ю. Д., Иомдин Л. Л.* (2010) Конструкция типа негде спать: синтаксис, семантика, лексикография // Апресян Ю. Д. и др. Теоретические проблемы русского синтаксиса: Взаимодействие грамматики и словаря, ЯСК, М. С. 59–113.

2. *Евтюхин В. Б.* (2008) Местоимение // Богданов С. И., Воейкова М. Д., Князев Ю. П. и др. Морфология современного русского языка. СПб.
3. *Гайнутдинова А. Ф.* (2012) Возвратные местоимения в современном русском языке. Автореф. … канд. филол наук. Казань.
4. *Егорова И. Н.* (1967) Позиционные эквиваленты слова в составе предложения (к изучению вариативных синтаксических рядов) // Русский язык. Грамматические исследования. М.: Наука. С. 78–95.
5. *Кронгауз М. А.* (1984) Тип референции именных групп с местоимениями все, всякий и каждый // Семиотика и информатика. Вып. 23. М., сс. 107–123.
6. *Кустова Г. И.* (2015) Прагматические факторы в редуцированных конструкциях: квазирелятивы типа что хочешь // Вестник Новосибирского государственного педагогического университета. № 5. С. 122–133.
7. *Лингвистика конструкций.* (2010) Под ред. Е. В. Рахилиной. М.
8. *Откупщикова М. И.* (1984) Местоимения современного русского языка в структурно-семантическом аспекте. Л.
9. *Падучева Е. В.* (1985) Высказывание и его соотнесенность с действительностью. М.
10. *Падучева Е. В.* (2013) Синтаксические типы общеотрицательных предложений // Падучева Е. В. Русское отрицательное предложение. М.: ЯСК, 2013. 304 с.
11. *Падучева Е. В.* (2015) Нереферентные местоимения на -нибудь // Русская корпусная грамматика. http://rusgram.ru
12. *Татевосов С. Г.* (2002) Семантика составляющих именной группы: кванторные слова. М.
13. *Тестелец Я. Г.* (2001) Введение в общий синтаксис. М.
14. *Тестелец Я. Г., Былинина Е. Г.* (2005) О некоторых конструкциях со значением неопределенных местоимений в русском языке: амальгамы и квазирелятивы / ИППИ РАН. Семинар «Теоретическая семантика». 15.04.2005 http://www.rsuh.ru/binary/1787534_99.1322270635.82662.pdf
15. *Шведова Н. Ю.* (1998) Местоимение и смысл: Класс русских местоимений и открываемые ими смысловые пространства. М.
16. *Шелякин А. Д.* (1986) Русские местоимения: Значение, грамматические формы, употребление. Тарту.
17. *Шмелев А. Д.* (2002) Русский язык и внеязыковая действительность. М.: ЯСК.
18. *Bošković Ž.* (2001) On the interpretation of multiple questions // Linguistic Variation Yearbook. Edited by Pierre Pica. Paris, John Benjamins Publishing Company, 270 pp.
19. *Bošković, Ž.* (2002) On multiple wh-fronting, On multiple wh-fronting. *Linguistic Inquiry,* 33, pp. 351–383.
20. *Dayal V.* (2005) “Multiple Wh Questions”, in M. Everaert and H. van Riemsdijk, Syntax Companion 3, Case #44. Blackwell Publishers, pp. 275–326
21. *Haspelmath M.* (1997) Indefinite Pronouns. — Oxford: Oxford University Press, 364 p.
22. *Kotek H.* (2014) Composing Questions. Massachusetts institute of technology.

References

1. *Apresjan Yu. D., Iomdin L. L.* (2010) Construction negde spat': syntax, semantics, lexicography, Apresjan Yu. D. et al. (2010) Theoretical problems of Russian syntax: The interaction of grammar and vocabulary, LSC, Moscow.
2. *Bošković Ž.* (2001) On the interpretation of multiple questions // Linguistic Variation Yearbook. Edited by Pierre Pica. Paris, John Benjamins Publishing Company, 270 pp.
3. *Bošković, Ž.* (2002) On multiple wh-fronting, On multiple wh-fronting. *Linguistic Inquiry,* 33, pp. 351–383.
4. *Construction linguistics* (2010) Ekaterina V. Rakhilina (ed.). Moscow.
5. *Dayal V.* (2005) "Multiple Wh Questions", in M. Everaert and H. van Riemsdijk, Syntax Companion 3, Case #44. Blackwell Publishers, pp. 275–326
6. *Egorova I. N.* (1967) Positional equivalents of words in sentence (to analysis of syntactic variable series) // The Russian. Grammatical studies. Moscow: Nauka Publ., pp. 78–95.
7. *Evtyukhin V. B.* (2008) Pronoum, Bogdanov S. I., Voeikova M. D., Knyazev Yu. P. et al. Morphology of modern Russian. St. Petersburg.
8. *Gaynutdinova A. F.* (2012) Reflexive pronouns in modern Russian. Kazan'.
9. *Haspelmath M.* (1997) Indefinite Pronouns. — Oxford: Oxford University Press, 364 p.
10. *Kotek H.* (2014) Composing Questions. Massachusetts institute of technology.
11. *Kronhauz M. A.* (1984) Reference type of noun phrases with pronouns vse, vsyakiy, kazhdy, Semiotics and Informatics, issue 23. Moscow, pp. 107–123.
12. *Kustova G. I.* (2015) Pragmatic factors in reduced construction: free relatives of hochesh series. Bulletin of NSPU, № 5, pp. 122–133.
13. *Otkupshchikova M. I.* (1984) Pronouns of modern Russian in structural and semantic aspects. Leningrad.
14. *Paducheva E. V.* (1985) Utterance and its relation to reality: Referential aspects of semantics of pronouns. Moscow: Science. 271 p.
15. *Paducheva E. V.* (2013) The Scope of the Negation. Presumption. Suspension of Assertion // E. V. Paducheva. Russian negative sentence. Moscow: Languages of Slavic Culture Publ., pp. 21–40.
16. *Paducheva E. V.* (2015) Non-referential pronouns with -nibud'// Russian corpus grammar. http://rusgram.ru
17. *Shelyakin A. D.* (1986) Russian Pronouns: meaning, forms, use. Tartu.
18. *Shmelev A. D.* (2002) Russian and extra-linguistic reality. Moscow.
19. *Shvedova N. Yu.* (1998) The pronoun and sense. Moscow.
20. *Tatevosov S. G.* (2002) Semantics of constituents of noun phrase: quantifier words. Moscow: IMLI RAS. 239 p.
21. *Testelets Ya. G.* (2001) Introduction to the general syntax. Moscow
22. *Testelets Ya, Bylinina E. G.* (2005) On some indefinite pronouns constructions: amalgams and free elatives. Report of the workshop "Theoretical semantics" under the direction of Ju. D. Apresjan. Moscow. 15.04.2005 http://www.rsuh.ru/binary/1787534_99.1322270635.82662.pdf

ЛЕКСИКАЛИЗОВАННАЯ ПРОСОДИЯ И ПОЛИСЕМИЯ ДИСКУРСИВНЫХ СЛОВ

Левонтина И. Б. (irina.levontina@mail.ru)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Работа посвящена ряду многозначных русских частиц и проблеме лексиколизованной просодии. Лексикализация просодии в русском языке привлекла внимание лингвистов около тридцати лет назад. При этом исследование данного явления ограничивается обычно изучением случаев лексикализации места и типа фразового ударения. Однако применительно к дискурсивным частицам оказывается, что имеются интересные случаи лексикализации не только фразового акцента, но и самого интонационного контура. В работе обсуждаются два таких случая. Во-первых, это частица *–то*, которая в разных своих значениях требует совершенно разного интонационного оформления фразы. Так, во фразах *Всё-то вы знаете; Уж бледная-то бледная; Где-то он бродит?* интонация резко отличается от интонации фраз с *то-* в большинстве других значений. Во-вторых, это частица *вот*, которая может использоваться не только как указательная частица, но и, например, как ксенопоказатель (маркер пересказа или цитирования); ср. *Пристала: вооот, что у тебя за юбка*. В этом последнем случае требуется весьма специфический интонационный контур фразы, так что данное значение частицы *вот* вообще плохо представлено в письменной речи.

Ключевые слова: лексикализованная просодия, полисемия, дискурсивные частицы, русский язык

LEXICALIZED PROSODY AND THE POLYSEMY OF DISCOURSE MARKERS[1]

Levontina I. B. (irina.levontina@mail.ru)

Russian Language (Vinogradov) Institute, Moscow, Russia

[1] Исследование выполнено за счет гранта Российского научного фонда (проект № 16-18-02054). Автор выражает глубокую признательность Т. Е. Янко за неоценимую помощь в анализе звучащей речи. Автор благодарит анонимных рецензентов «Диалога» за содержательные замечания.

The paper is devoted to some polysemous Russian particles and the problem of lexicalized prosody. The phenomenon of lexicalized prosody in Russian drew the linguists' attention about 30 years ago. The investigation is usually confined to phrasal stress as the most frequently lexicalized and therefore the most lexicographically interesting prosodic pattern. However, as far as discourse particles are concerned, not only phrasal stress but also intonation pattern is of greatest interest. Two Russian discourse particles will be discussed. One of them is the particle *-to*. Its different usages imply very different prosodic patterns. The other one is *vot* which can be used not only as a demonstrative particle, but also as a xenomarker (quotation marker), which requires a specific prosody.

Key words: lexicalized prosody, polysemy, discourse particles, Russian

1. The phenomenon of lexicalized prosody

The phenomenon of lexicalized prosody in Russian drew the linguists' attention about 30 years ago. See [Apresyan: 1980: 51; Nikolaeva 1985: 122; Pavlova 1987:8; Bulygina, Shmelev 1987; Zaliznyak Anna 1994, Boguslavsky 1996: 255, etc.]. A very nice overview of different approaches to this problem can be found in [Kobozeva, Zaharov 2004].

Mainly this phenomenon is connected with the communicative structure of an utterance, the topic/focus opposition.

Thus, in ADR (Активный словарь русского языка) lexicalized prosody is treated the following way: "Prosody comprises a wide range of phenomena, out of which ADR is concerned primarily with those that pertain to phrasal stress as the most frequently lexicalized and therefore the most lexicographically interesting prosodic pattern.

Lexicalized phrasal stress is always in some way tied up to the communicative structure of the sentence. There are two groups of prosodic phenomena that are reflected in ADR: a) prosodic syntagmatics; b) prosodic pardigmatics.

a) Prosodic syntagmatics. Certain lexemes, while themselves not prosodically marked, do, at the same time, require prosodic accentuation of the words they are syntactically connected with. Thus, the particle *čto kasaetsja* 'as for X,' 'what concerns X' marks the NP to the right of it, on which it is syntactically dependent, as the contrastive rheme of the sentence, and b) Prosodic paradigmatics. Prosodic paradigmatics deals with prosodic accentuation as a marker of differences among various lexemes of the same word or different usages of the same lexeme. Prosodic accentuation tends to mark the following categories of meanings: 1) negation; 2) quantification; 3) modalities of desire, necessity and possibility; 4) evaluation; 5) facts and opinions. The presence of one or more of those meanings in the semantics of a lexical item allows one to form expectations concerning its prosodic properties.

This phenomenon can be illustrated with two different lexemes of the word *pozdno* 'late.'

Pozdno 1 'late 1' means 'at a late time' and can either bear phrasal stress or be prosodically unmarked, whereas *pozdno 3* which means 'too late for doing X' and

combines the components of quantification, lost possibility, and evaluation, always bears the main phrasal stress" [Apresjan V. 2011].

As we see, the main attention is paid to phrasal stress. However it should be mentioned that as far as discourse particles are concerned, not only phrasal stress but also intonation pattern is of greatest interest [see also Kodzasov 1996]. I will provide two examples to illustrate my statement.

2. *–TO*

2.1. In Russian two homonymous particles *-to* are present:

1) a particle forming indefinite pronouns: *кто-то* somebody, *какой-то* some, *почему-то* for some reason, etc.;
2) a discourse particle with a number of meanings [see Shimchuk, Shchur 1999].

-To as a discourse particle is discussed in many articles, most thoroughly in [Bonnot 1990, 1991]. According to Bonnot, *-to* is mostly a topicalizing particle—"la particule de thematisation".

However it has some lexemes implying a very unusual intonation pattern.

Let us compare the intonation of the following phrases[2]:

(1a) *Всё-то сразу не трать!* Don't spend all you money at once!

(1b) *Всё-то я не потрачу!* I will not spend all my money at once!

(1c) *Всё-то вы знаете!* You know really everything!

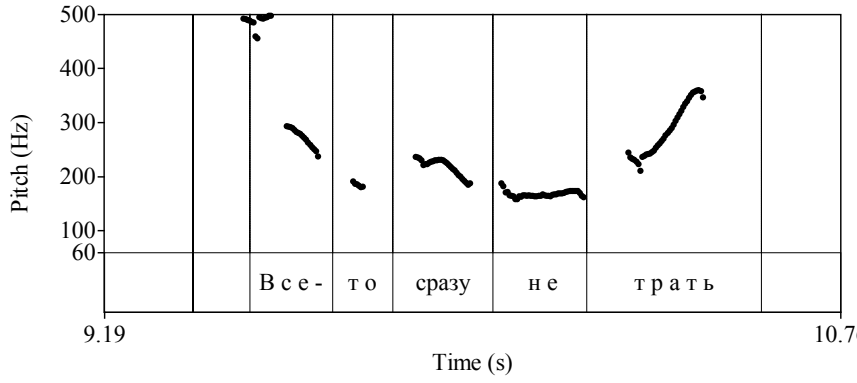


Fig. 1

[2] For speech analyzing the the Praat analyzer is used, see http://www.fon.hum.uva.nl/praat/. All phrases were read by one speaker (not by the author), female voice.

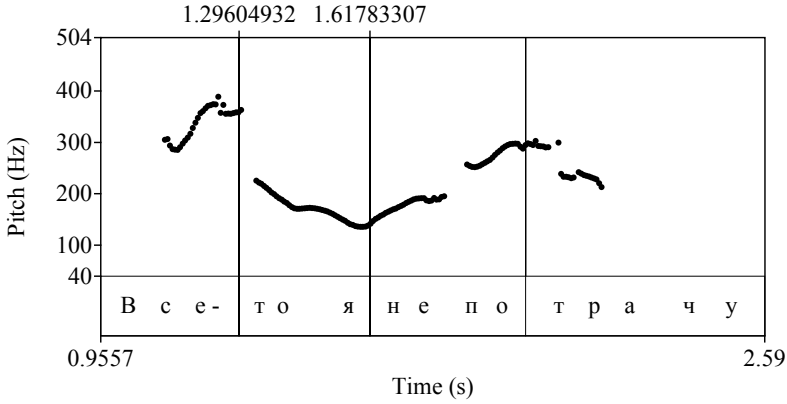


Fig. 2

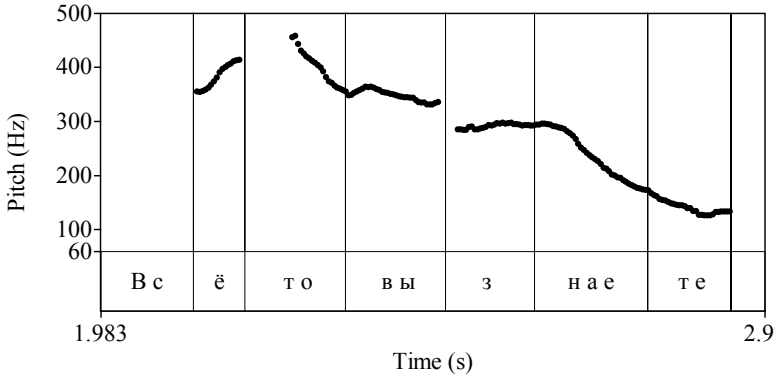


Fig. 3

Fig. 1 and Fig. 2 look different, because (1a) is an imperative sentence, which pressuposes a specific communicative structure, while in (1b) we see the regular topicalizing *–to*. And it is natural that here the so-called intonation construction 3, according to E. A. Bryzgunova's classification [see Bryzgunova 1977; Russian grammar 1980] is represented. This intonation pattern is typical of general questions, for example.

However, the intonation of (1c) differs greatly both from (1a) and (1b). The example (1c) and Fig. 3 demonstrate the so called intonation construction 5 as defined by Bryzgunova which is characteristic of the Russian exclamations [Russian grammar 1982: 116–117] (its frequency contour is specified by the two—the sentence-initial and the sentence-final—tonal peaks).

The respective meaning of the particle *–to* can be realized in the sentences with a certain generalization: either containing a word like *all, every, everything, always, never*, etc., or a sort of a row:

Ох, человек, человек, всех-то ты победил, всюду-то ты наводишь порядок, только среди самого человечества одни раздоры, скандалы, свары. [Виктор Розов. Удивление перед жизнью (1960–2000)]

Всегда-то он первый догадается! — прямо завидно. [Андрей Битов. Моя зависть (1965)]

Уж и вино отпустил! Можно сказать, что на чести. Попробовала я рюмочку, так и гвоздикой-то пахнет, и розаном пахнет, и еще чем-то. [А. Н. Островский. Бесприданница (1879)]

В биографии Галича, написанной Михаилом Ароновым и переизданной в прошлом году «Новым литературным обозрением», собрана целая россыпь цитат, полных плохо скрываемой зависти: и носки-то у него белые, и брюки трубочкой, и цепочка от часов золоченая. И на стол-то накрывают — все на фарфоре, и на бегах он пропадает, и ужинает в «Национале», и запел он, как писал Нагибин, «от тщеславной обиды» (Ю Сапрыкин http://kommersant.ru/doc/2315038?fb_action_ids=459096970874370&fb_action_types=og.likes&fb_source=other_multiline&action_object_map=%7B%22459096970874370%22%3A584859334907138%7D&action_type_map=%7B%2245909 6970874370%22%3A%22og.likes%22%7D&action_ref_map=%5B%5D)

Now let us compare intonation of similar phrases with and without –*to*:

(2a) *ВСЕГДА ты споришь!* and *Всегда ты СПОРИШЬ!*

(2b) *Всегда-то ты споришь!*

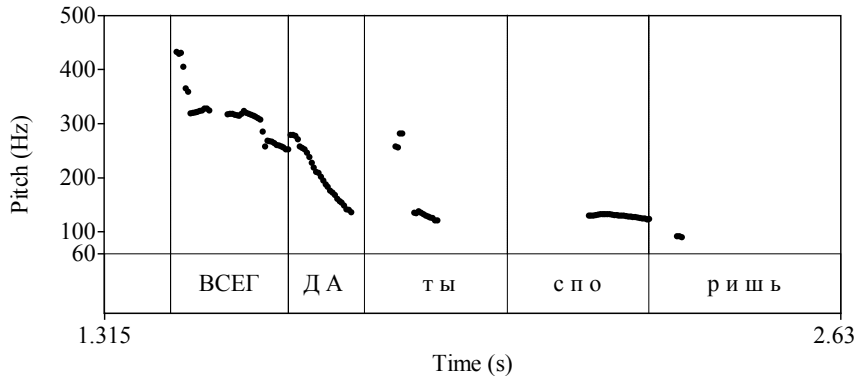


Fig. 4

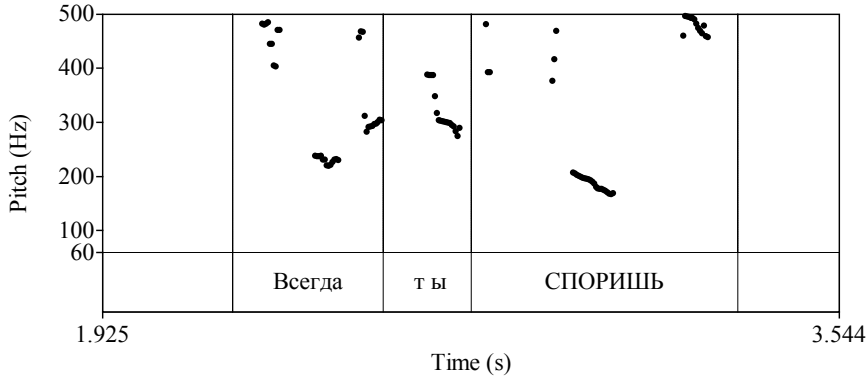


Fig. 5

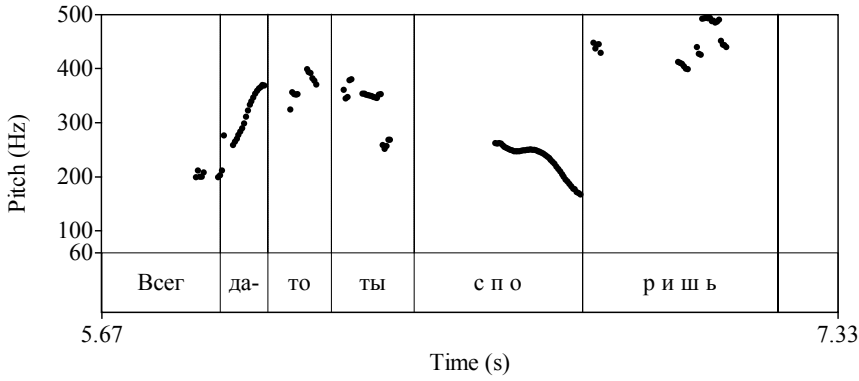


Fig. 6

The phrases without *–to* are pronounced with a typical rhematic accent (here so-called intonation construction 2), two variants differ by location of phrasal accent.

The phrase with *–to* from the point of view of intonation is very similar to our example (1c) *Всё-то вы знаете!* (intonation construction 5).

From the point of view of semantics the difference between phrases with and without *–to* is also clear.

Both phrases contain generalization (*You **always** object*).

But without *–to* it is just a statement, while with *–to* the speaker pretends that the general idea itself is well known and he or she wants just to illustrate it and express certain feelings towards the situation that cannot be changed.

2.2. There is one more lexeme of *–to* in combination with words like *who?, what?, where?, when?*, etc.

So, these combinations are homonymous to indefinite pronouns, but have different meaning and imply a specific intonation pattern. Let us compare:

(3a) *И он не с ↓теми ходит где-то.*

(3b) *Опять он где-то ↓ходит!*

(3c) *Он где-то ходит?*

Где-то has here the meaning 'somewhere', *-to* is prosodically unmarked.

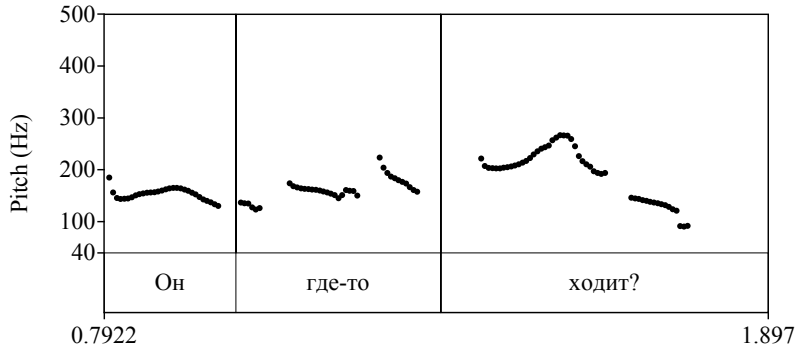


Fig. 7

(3d) *Где-то он ходит?* 'If I knew where he could be now!'

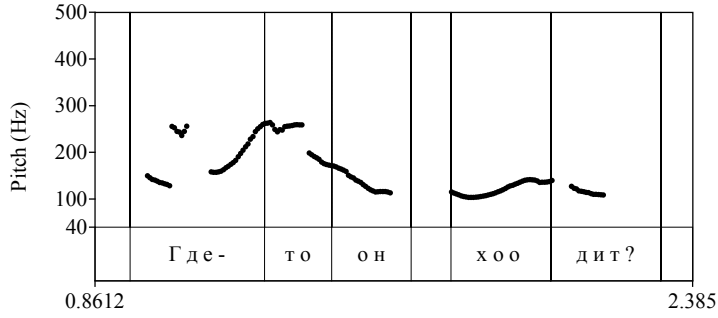


Fig. 8

Here a very interesting type of prosody is represented. It is typical of a specific illocution of mirroring the process of dreaming, or recollecting. It is marked by a rise on the tonic syllables which are followed by level, or slightly falling post-tonics. The tonic and post-tonics are substantially prolonged; see [Yanko 2008: 109]. The meaning of this *–to* can be described the following way: 'The answer to this question can hardly be found, but I am trying to imagine what it could be'.

2.3. One more lexeme of *–to* implying similar intonation pattern is worth mentioning. See the following examples:

(4) *Уж бледная-то бледная;*

Зашёл разговор о лошадях, и Печорин начал расхваливать лошадь Казбича: уж такая-то она резвая, красивая, словно серна, — ну, просто, по его словам, этакой и в целом мире нет [М. Ю. Лермонтов. Герой нашего времени (1839–1841)]*;*

Порадовались черемухе, все в нее головами нюхали, самая-то весна [И. С. Шмелев. Лето Господне (1927–1944)].

Here the particle stresses the idea of extremely high level or extent:

Fig. 9

So we tried to demonstrate that lexicalized prosody is not always confined to phrasal stress. We see that some lexemes of the discourse particle *–to* imply certain intonation patterns.

3. *ВОТ*

The particle *вот* is also polysemous[3]. Its basic meaning is demonstrative (*Вот моя деревня*). Some of the lexemes are stressed, some are not, some of them are prosodically unmarked. There is an interesting function of this particle as "xenomarker".

---

[3] **ВОТ**, ЧАСТ.
вот 1.1 'говорящий что-то показывает адресату': *Вот моя квартира*.
вот 1.2 'говорящий сообщает адресату ответ на какой-л. вопрос': *Вот что мы сделаем — мы ему обо всем расскажем сами.*
вот 2 наррат. 'говорящий предлагает адресату представить себе что-л.': *Вот пришел он к ней и говорит: "Выходи за меня замуж".*
вот 3.1 'говорящий сообщает, что время события близко к моменту речи': *Вот только что <сию минуту> я его видел, и уже опять он куда-то исчез.*

This function is usually ascribed to the particles *мол, дескать, де,* as well as *якобы* and *грит (гыт)*. Most of them are etymologically connected with verba dicendi. Since the speaker uses xenomarkers to distance himself or herself from another person's stand, these words quite often pragmatically imply a valuation, most often a negative one, of the reported speech. Their function, according to N. D. Arutyunova, consists in "marking Somebody Else's presence" («маркировать присутствие Другого») [Arutyunova 2000: 448]. It turns out, however, that the repertoire of means used as markers of quotation or retelling is much broader than it is generally admitted. Thus, the words *ах, вот, так и так*; the construction with imperative reduplication and the conjunction да (*Привязалась: расскажи да расскажи*), specific intonations of retelling, and some other phenomena can take over the same function [Levontina 2011].

Вот as xenomarker was discovered in [Podlesskaya, Kibrik 2009], namely as a means of "expressing threat and condemnation" in reported speech, cf.:

\тоже на меня '' /посмотрела , «/\Вот! Я тебя /–в-выгоню-у и-из-зз ··(0.2) этой из ш= ··(0.2) ''из ··(0.2) '' /школы!»
She looked at me too: "I'll throw you out of this school!".

See also:

Я стала говорить дома: вот, Наташа просила у меня прощения, я не простила ее...

вот 3.2 'говорящий подчеркивает, что какое-то событие произошло незадолго до момента речи': *Вот и лето пришло.*
вот 3.3 разг. 'говорящий сообщает актуальную информацию о себе:' — *Ты откуда?* — *Да вот приехал отца навестить.*
вот 4 'говорящий подчеркивает, что желательное событие произойдет сразу после другого события': *Вот сделаю уроки и пойду гулять.*
вот 5 'говорящий угрожает адресату': *Вот скажу маме, как ты меня обзываешь, она тебе задаст!*
вот 6.1 'говорящий выделяет кого-л. или что-л.': *Вот вы, к примеру, что об этом думаете?*
вот 6.2 'говорящий подчеркивает, что он переходит от общих утверждений к конкретным примерам': *Так вести себя некрасиво. Вот скажи, разве тебе было бы приятно, если бы тебя стали дразнить плаксой?*
вот 7 'поэтому': *Он сломал велосипед, вот пусть теперь и чинит его.*
вот 8.1 разг. 'говорящий дает эмоциональную оценку какого-л. объекта или ситуации': *Вот сумасшедший!*
вот 8.2 разг. 'говорящий подчеркивает несогласие с адресатом': *А вот и ошибаешься.*
вот 8.3 'говорящий выражает желание, чтобы имела место какая-то ситуация': *Вот бы оказаться сейчас на берегу моря.*
вот 8.4 разг. 'говорящий подчеркивает свое одобрение действий адресата': *Вот, правильно.*
вот 9 разг. 'говорящий не знает, что сказать дальше': *Меня зовут Ваня. Вот.*
вот 10 разг. 'говорящий передает чужие слова': *Мне говорят, вот, я плохая мать.*
(ADR, lexical entry was written by T. Krylova)

I started telling at home—well, Natasha asked me to forgive her, and I didn't…
[Н. Горланова, Метаморфозы].

Interestingly enough, the meaning of *вот* is not confined to threat or condemnation. Compare the following examples:

А она сидит и ноет: «Вот, я такая несчастная…»
And she is sitting around, whimpering: "Oh, I'm so miserable…";

Он расхвастался: «Вот, я самый крутой»
He started bragging: "Yeah, I'm such a cool guy";

Привязалась: «Вот, как тебе не стыдно, что у тебя за юбка»
She kept intruding on me: "Hey, what kind of a skirt is that? Shame on you!";

А он все обещает: «Вот, деньги будут со дня на день, все отдам»
And he keeps giving promises: "The money will be there any day, I will return everything I owe";

Ну и что же, что она первая позвонила? А ты бы ей сказал: «Вот, я сам собирался тебе позвонить, поздравить»
Yes, she called first, so what? You should have said—I was planning to call you with the greetings myself.

It should be stressed that this lexeme demands certain intonation.

In [Yanko 2008: 109] an intonation pattern described is called "the intonation of mental activity", i.e. situations of remembering, perplexity, sinking into daydreams, and also reported speech (*Тетя сказала, надо чего-то там уко-олы делать* [Aunty said we should make injections…]). This intonation pattern is described as follows: «Соответствующий акцент характеризуется подъемом тона и существенным удлинением ударного слога акцентоносителя ремы. Вся заударная область ровная (иногда с небольшим естественным падением)» [*A rheme of dreaming or recollecting*: a rise on the tonic syllables followed by level or slightly falling post-tonics accompanied by lengthening the tonic and post-tonic syllables.]. Yanko notes, that by retelling a speaker does not copy somebody else's intonation, but arranges his or her utterance with a specific "remembering" prosody.

However, although the prosody of retelling and remembering has much in common, these two intonation patterns are somewhat different. First of all, retelling is often affective and emotive, and the prosody in this case is emphatic, which is hardly possible in the case of remembering, perplexity, or sinking into daydreams. Secondly, by retelling, the phrase is often split into minor segments as against original speech.

Cf.:

— И что он ответил?
— Да что ответил! «/\Маама не разре/\шаает»

— And what did he say?
— What could he say? Mommy won't let me.

From the point of view of intonation, reported speech often turns to sounding rhythmical, pronounced with seriate tone rises and falls, similar to "listing" intonation:

А он мне и говорит: «/\Воот, /\деевушка, какая вы кра/\сиивая, как вас зо/\вут, а пой/\деемте, погу/\ляем, а /\даайте теле/\фоончик»
And he tells me: Hey lady, you're so pretty, what's your name, let's go for a walk, please give me your phone number.

For phrases with *вот* as xenomarker this type of intonation is obligatory: the particle only cannot express this meaning without this pattern, it implies certain prosodic outline of the whole phrase. That's why it occurs mostly in oral speech.

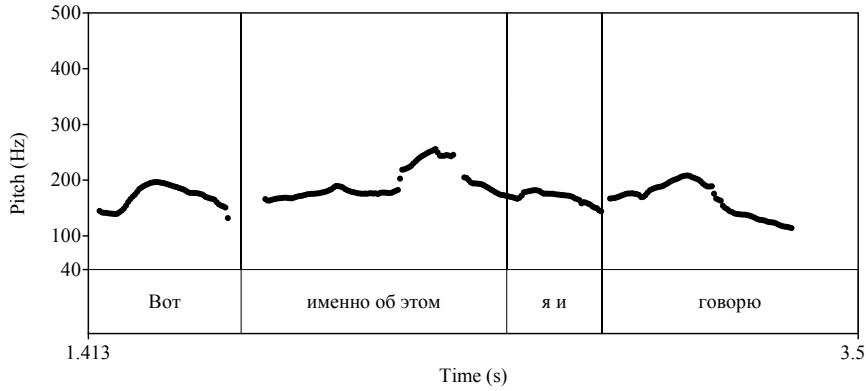


Fig. 10

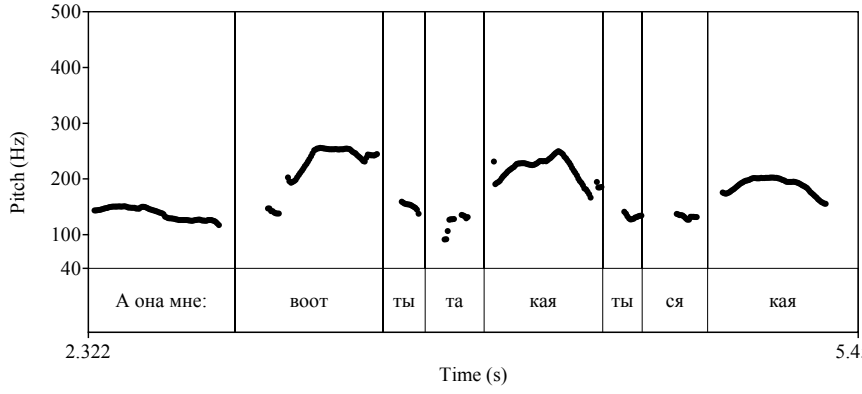


Fig. 11

4. Conclusion

Such cases as our *–mo* and *вот* are much more difficult for lexicographical presentation than the cases of lexicalized phrasal stress. As was mentioned earlier, in ADR such features of lexemes are so far not being fixed. I hope, however, that eventually this problem will be solved either with the help of short and clear verbal descriptions, or with signs like /\, /-\, or similar. For the future digital version of ADR audio-illustrations will be helpful in such cases [see Kobozeva, Zaharov 2004].

References

1. *ADR* (2014), Apresyan V. Yu., Yu. D. Apresyan, E. Eh. Babaeva, O. Yu. Boguslavskaya, I. V. Galaktionova, M. Ya. Glovinskaya, B. L. Iomdin, T. V. Krylova, I. B. Levontina, A. I. Lopuhina, A. V. Ptencova, A. V. Sannikov, E. V. Uryson. Active Dictionary of Russian [Aktivnyj slovar› russkogo yazyka]. Vv. 1–2. Ed. by Yu. D. Apresyan.—M.: Yazyki slavyanskoj kul›tury, 2014.
2. *Apresjan V.* (2011), Active Dictionary of Russian: Theory and Practice, Proc. 5th Int. Conference on Meaning-Text Theory. Barcelona, September 8–9, 2011. Ed. by Igor Boguslavsky and Leo Wanner. Barcelona: University PompeuFabra, 2011. http://meaningtext.net/mtt2011/proceedings/papers/Apresjan.pdf
3. *Apresyan Yu. D.* (1980), Types of information for the Surface syntax component of the Meaning-Text Model [Tipy informacii dlya poverhnostno-semanticheskogo komponenta modeli «Smysl ↔ Tekst»]. Wien: Wiener Slawistischer Almanach.
4. *Arutyunova N. D.* (2000), Xenomarkers *de, deskat', mol* [Pokazateli chuzhoj rechi *de, deskat', mol*], Yazyk o yazyke. Ed. by N. D. Arutyunova. Moscow. Pp. 437–452.
5. *Boguslavsky I. M.,* (1996), Scope of lexical units [Sfera dejstviya leksicheskih edinic]. Moscow.
6. *Bryzgunova, E. A.* (1977), Sounds and intonations of Russian speech [Zvuki i intonacii russkoj rechi]. Moscow, Russian language[Russkij yazyk].
7. *Bonnot Chr.* (1990), The particle *–to* and means of polemic [La particule *-to* et la polemique cachee en russe moderne: A propos du statut enonciatif du theme] // Rev. des etudes slaves.—P.—T. 62, fasc 1/2.—Pp. 67–75.
8. *Bonnot Ch.* (1991), The particle *–to* in modern Russian [La particule de thematisation *-to* en russe modern] // Rev. des etudes slaves.—P.—T. 63, fasc 4.—P. 853–861.
9. *Bulygina T. V., Shmelev A. D.* (1987), Semantics of the Russian particles *razve* and *neuzheli* [O semantike chastic *razve* i *neuzheli*], NTI № 10. Pp. 21–25.
10. *Kobozeva I. M., Zaharov L. M.* (2004), Why do we need an audial dictionary of Russian discourse words [Dlya chego nuzhen zvuchashchij slovar' diskursivnyh slov russkogo yazyka] // Trudy mezhdunarodnoj konferencii Dialog'2004 «Komp'yuternaya lingvistika i intellektual'nye tekhnologii». Moscow, Science [Nauka]. http://www.dialog-21.ru/Archive/2004/Kobozeva.pdf

11. *Kodzasov S. V.* (1996), Semantic and phonetic splitting of Russian particles and prosodic unformation in the dictionary [Semantiko-foneticheskoe rasshcheplenie russkih chastic i prosodicheskaya informaciya v slovare], Dictionary. Grammar. Text [Slovar›. Grammatika. Tekst]. Mjscow. Pp. 97–112.
12. *Levontina I.* (2011), Xenomarkers in Russian, Proc. 5th International Conference on Meaning-Text Theory. Barcelona, September 8–9, 2011. Ed. by Igor Boguslavsky and Leo Wanner. Barcelona: University Pompeu-Fabra http://meaningtext.net/mtt2011/proceedings/papers/Levontina.pdf
13. *Nikolaeva T. M.* (1985), Functions of particles in an utterance in Slavic languages [Funkcii chastic v vyskazyvanii. Na materiale slavyanskih yazykov]. Moscow.
14. *Pavlova A. V.* (1987), Accent structure of an utterance in connection with lexical semantics [Akcentnaya struktura vyskazyvaniya v ee svyazyah s leksicheskoj semantikoj]. Avtoreferat dis…. kand. filol.nauk. Leningrad.
15. *Podlesskaya V. I., Kibrik A. A.* (2009), Discourse markers in the structure of oral narrative [Diskursivnye markery v strukture ustnogo rasskaza: opyt korpusnogo issledovaniya] // www.dialog-21.ru/digests/dialog2009/materials/pdf/60.pdf
16. *Russian grammar [Russkaya grammatika],* (1980), Moscow, Science [Nauka.].
17. *Shimchuk, Eh. G.; Shchur, M. G.* (1999), Dictionary of Russian particles [Slovar' russkih chastic]. Frankfurt am Main—Berlin—Bern—Bruxelles—New York—Wien. Peter Lang—Europäische Verlag der Wissenschaften.
18. *Yanko T.* (2008), Intonational strategies of the Russian speech from a contrastive perspective [Intonatsionnye strategii russkoj rechi v soposnaviteljnom aspekte]. Moscow.
19. *Zaliznyak Anna A.* (1994), The celebration of life is passing by (on the ambiguity of some Russian words) [Prazdnik zhizni prohodit mimo (Zametki o neodnoznachnosti nekotoryh russkih slov)]. // Wiener Slavistischer Almanach, B. 34, p. 261–278.

COMPARISON OF MELODIC PORTRAITS OF ENGLISH AND RUSSIAN DIALOGIC PHRASES

Lobanov B. M. (Lobanov@newman.bas-net.by)

United Institute of Informatics Problems NAS Belarus, Minsk

This study is an extension of the author's works, presented at the "Dialogue 2014 and Dialogue 2015" conferences. According to the concept of universal melodic portrait (UMP), a phrase intonation can be described as a sequence of UMPs of accentual units (AUs) that make up the phrase. The present paper describes the results of pilot studies where melodic portraits for English and Russian language phrases were compared. The examined phrases were derived from simple situational dialogues and were spoken by native English and Russian speakers. The study was restricted only to phrases with a one-accent unit structure representing the three main types of phrase intonations: affirmative statements, special questions and general questions.

The described UMP model allows to investigate tonal differences within languages by applying precise quantitative assessments. The method can be used effectively for solving problems of language interference. Moreover, the UMP model could potentially find an effective application in foreign language studies. Using the appropriate software that realizes the described stages of UMP construction, a learner could be able to visually compare an intonation of the pronounced phrase with its target intonation portrait and work to eliminate a foreign accent by proper training.

Key words: intonation patterns, melodic portrait, synthesis and analysis of intonation, English and Russian intonations, English and Russian as second languages, TTS synthesis

СРАВНЕНИЕ МЕЛОДИЧЕСКИХ ПОРТРЕТОВ АНГЛИЙСКИХ И РУССКИХ ФРАЗ ДИАЛОГОВОЙ РЕЧИ

Лобанов Б. М. (Lobanov@newman.bas-net.by)

Объединённый институт проблем информатики
НАН Беларуси, Минск, Беларусь

Ключевые слова: интонационные конструкции, мелодический портрет, синтез и анализ интонации, английская и русская интонации, английский и русский как иностранный

Introduction

The present work is a follow up study to the previously introduced model of universal melodic portraits (UMP) of accentual units[1] (AU) for representation of phrase intonations in TTS synthesis [Lobanov et al, 2006]. According to this model, a phrase is represented by one or more of AUs. Each unit, in turn, can be composed of one or more phonetic word. If there is more than one word in an AU, than only one word bears the main stress while other words carry a partial stress. Each AU consists of ***pre-nucleus*** (all phonemes preceding the main stressed vowel), ***nucleus*** (the main stressed vowel) and ***post-nucleus*** (all phonemes following the stressed vowel). The UMP model assumes that topological features of melodic AU for particular type of intonation do not depend on a number or quality of phonemic content of a pre-nucleus, nucleus or post-nucleus, nor on the fundamental frequency range specific for a given speaker.

The UMP model allows to represent intonation constructs as a set of melodic patterns in normalized space {***Time—Frequency***}.

Time normalization is performed by bringing pre-nucleus, nucleus and post-nucleus elements of AU to standard time lengths. This sort of normalization levels out the differences in melodic contours caused by the number of words and phonemes in an AU.

For fundamental frequency normalization $F_{0\,min}$ and $F_{0\,max}$ are determined within the ensemble of melodic contours produced by a certain speaker. This sort of normalization cancels out the differences of melodic contours caused by speakers voice register and diapason.

The normalization is calculated by the formula

$$F_0^N = \frac{(F_0 - F_{0\,min})}{(F_{0\,max} - F_{0\,min})} \tag{1}$$

In certain cases it may be beneficial to use statistical normalization instead of (1)

$$F_0^N = \frac{(F_0 - M)}{\zeta} \tag{2},$$

where M is mathematical expectation, ζ is standard deviation. Note that M can be interpreted as a register and ζ—as a diapason of speaker's voice.

Therefore, the normalized space for UMP may be presented as a rectangle with axes (T_N, F_0^N) as schematically shown in **Figure 1**, while the interval [0–1/3] on the abscess T_N is a pre-nucleus, [1/3–2/3] is a nucleus, and [2/3–1] is a post-nucleus. The intervals on the ordinate F_0^N: [0–1/3]—low level, [1/3–2/3]—mid-level, [2/3–1]—high level.

[1] Accent Unit often referred to as Accent Group [Ogden et al, 2000]

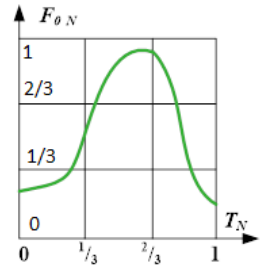


Figure 1. MPAU-representation

Figure 2 illustrates the results of time-frequency normalization of the example one-accent-unit phrases with affirmative intonations: *"It is no distance* ***at all****"* and *"It is only a couple of hundred* ***yards****"*.

The first phrase contains four phonetic words (underlined) and the second one—five. The last word in both phrases is accented (in bold font), and the nucleus is the stressed vowel in this word. **Figure 2** shows the intonograms of both phrases obtained with the PRAAT package (see: http://www.fon.hum.uva.nl/praat/). The figure demonstrates that phrases spoken by different speakers differ by 1.5 times in duration and 1.3 times in the maximum fundamental frequency. Despite these lexical and fundamental frequency differences, the final construction of UMPs for both phrases (the right-upper part of **Figure 2**) makes the similarity of melodic portraits evident.

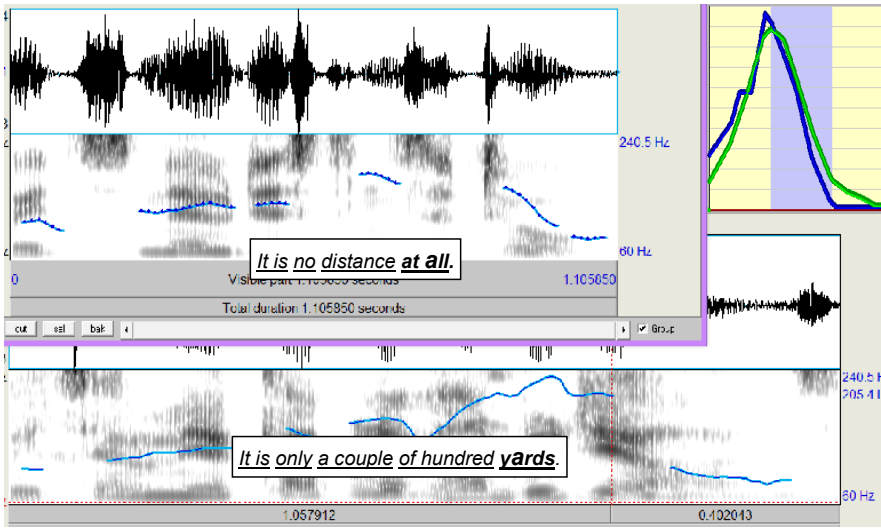


Figure 2. Illustration of time and frequency normalization

In the earlier work [Lobanov, 2014], the efficiency of suggested approach was verified by constructing UMPs for main intonation patterns of Russian speech: IP1—IP7. The subsequent study [Lobanov, 2015] demonstrated successful construction

of UMPs for compound narrative sentences in Russian. The present study provides pilot results for comparison of UMPs of English and Russian phrases for simple dialogue spoken by native English and Russian speakers.

The paper is laid out in the following way: the first paragraph describes the chosen texts and audio-material as well as the method of applying MPAU model to the analysis, the second paragraph shows the results of MPAU modeling and also the analysis and interpretation of the results obtained.

1. Method

The experiment was based on English texts and audio-files from the manual [Ockenden, 2005] which included:

- 44 everyday situations, each containing four dialogues in natural conversational English;
- All dialogues consist 1051 sentences, including 704 affirmative, 325 interrogative and 22 exclamatory sentences, spoken by certain number of male and female speakers;
- Situations relevant to those studying or travelling in England, including eating out, entertainment and travel, as well as more general functions such as greetings, complaining and apologizing.

In the present study we have restricted ourselves to three major types of phrase intonation—***Affirmative statements***, ***Special questions*** and ***General questions***. In addition, we restricted the study of intonations to the case of one-AU phrases (it is about 70 par cents of whole number of phrases). Other intonation types such as *Alternative questions, Tag questions, Commands, Exclamatory sentences, Direct address, Enumerating, Introductory phrases etc.* were not included in this study.

The comparison Russian language test material was based on direct translations of corresponding English phrases into Russian. The translated text was used to make Russian audio recordings that imitated normal conversation of two people with a standard Russian accent.

The composition of UMPs of both Russian and English phrases was performed with the aid of *PhonoClonator* and *IntoClonator* systems [Lobanov, 2014]. On the basis of a pre-marked text, the *PhonoClonator* system makes it possible to automatically segment each signal into phonemes and pitches (F0) and indicate positions of a nucleus for AU in a phrase.

Figure 3 shows the general view of the users interface of the *PhonoClonator* system for phrase processing.

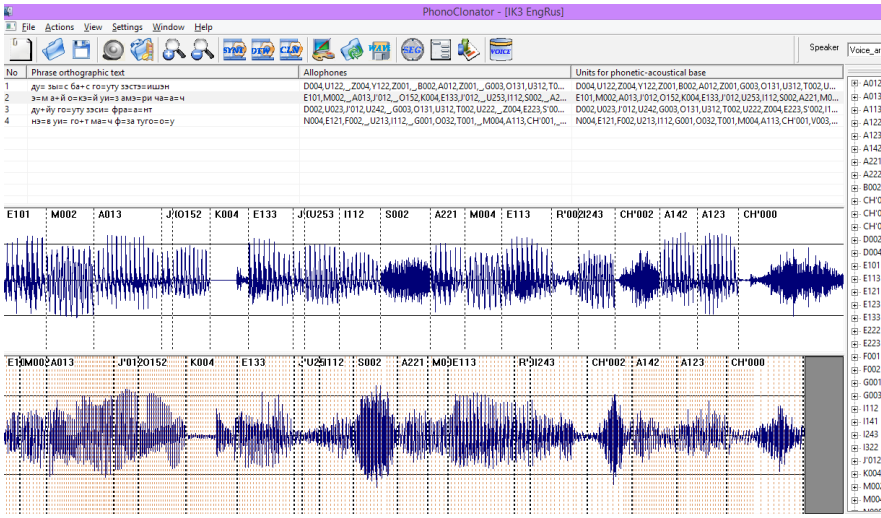


Fig. 3. PhonoClonator: the general view of the users interface

In the next step, the pre-marked audio-signals are fed into ***IntoClonator*** system that provides the boundaries of the nucleus, pre-nucleus and post-nucleus as well as melodic and intensity contours (Fig. 4). Minimum ($F_{0\,min}$) and maximum ($F_{0\,max}$) fundamental frequency values (F0) are determined automatically for the melodic contour of the phrase analyzed—*"Am I OK for St Marys Church?"*

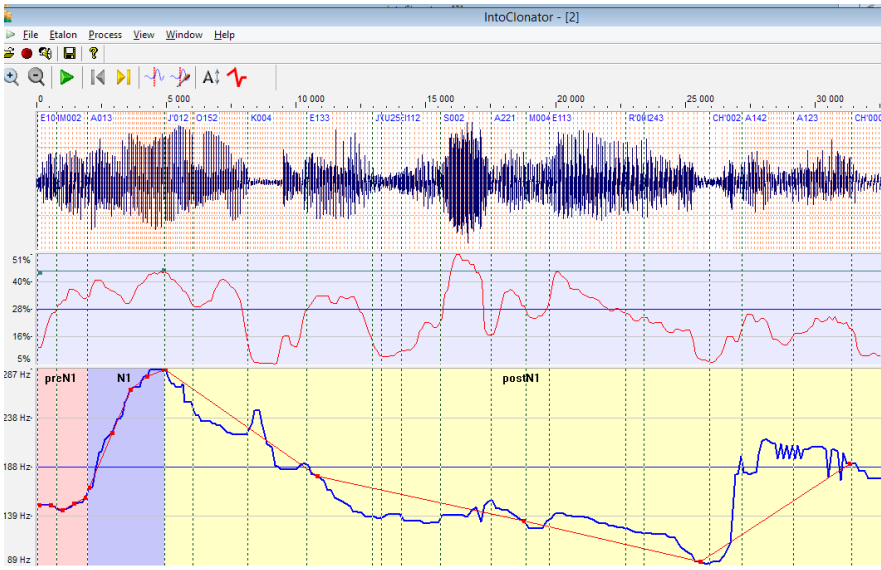


Fig. 4. IntoClonator: the general view of the users interface

Finally, ***ShapeEditor*** system makes it possible to use the information processed by ***IntoClonator*** system for composing melodic portraits of the analyzed phrase *"Am* ***I*** *OK for St Marys Church?"* in a normalized UMP-form described above (see: Figure 5).

Fig. 5. ShapeClonator: the general view of the users interface

2. Results

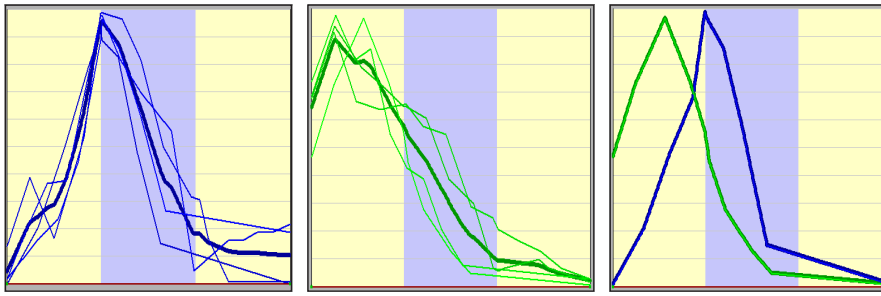
Here, we present the results of comparisons of melodic portraits of English and Russian phrases chosen from sample dialogues on the principle of being well-pronounced examples of the three evaluated types of intonation contours: affirmative statements, special questions and general questions.

For affirmative statements we used English and Russian phrases an example of which is listed in Table 1. The phrases were spoken by different speakers. The analyzed one-accent-unit phrases are italicized. The word that carries the main accent is printed in a bold type with its stressed vowel (nucleus) underlined. All syllables to the left of the nucleus make up a pre-nucleus and those to the right—a post-nucleus.

Table 1. English and Russian phrases spoken with affirmative intonation of statements

English	Russian
- Is it far? - *It is only about five minutes* ***walk****.*	- Далеко ли это? - *Это всего в пяти минутах* ***ходьбы****.*
- Will it take me long to get there? - *It is no distance* ***at all****.*	- Долго ли мне придётся идти? - *Это вообще* ***не расстояние****.*
- Should I take a bus? - *You can walk it in under five* ***minutes****.*	- Мне нужно подождать автобуса? - *Вы сможете дойти за пять* ***минут****.*
- Is it too far to walk? - *It is only a couple of hundred* ***yards****.*	- Долго ли придётся идти пешком? - *Это всего в паре сотен* ***шагов****.*

Figure 6, as well as following **Figures 7** and **8**, show melodic portrait curves obtained with the use of computational approaches described in the **Introduction**. In Figure 6 (a), thin blue lines reflect the melodic portraits of four English phrases and the bold line reflects the averaged UMP. The UMP is represented along the X-axis by the succession of three time normalized stretches—pre-nucleus, nucleus, post-nucleus, with normalized fundamental frequency relative to the phrase maximum and minimum along the Y-axis. Similarly, in **Figure 6 (b)** green lines show tone curves for the Russian phrases, and in **Figure 6 (c)** shows superimposed typical intonation contours for English and Russian affirmative statements.



(a) English (b) Russian (c) English & Russian

Figure 6. UMPs for English and Russian one-accent-unit phrases (affirmative statements)

The comparison of English and Russian affirmative statement melodic portraits in **Figure 6 (c)** allows to establish the following differences:

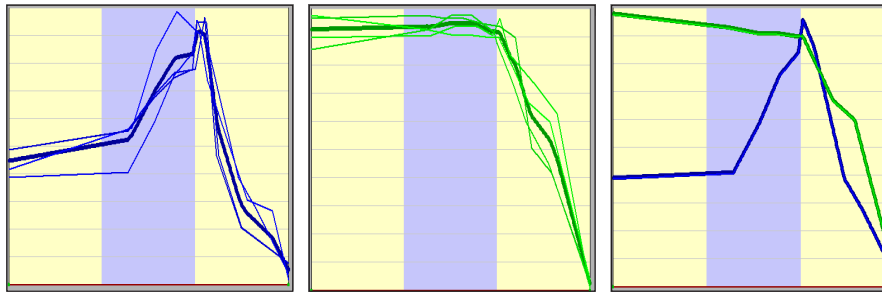
- the most changes are found in pre-nucleus and nucleus regions;
- in the pre-nucleus region, maximum of the Russian UMP curve falls closer to the middle of the region, whereas the English UMP curve peaks at the end;
- in the nucleus region, the English UMP curve is characterized by a sharper decline in comparison to the Russian UMP curve.
- in the post-nucleus region, both Russian and English MP curves show an identical low level steady decline.

Next, for the study of intonation characteristics of special questions, we used the example English and Russian phrases listed in **Table 2.** The content representation and mark up in **Table 2** is similar to **Table 1**.

Table 2. English and Russian phrases spoken with the intonation of special questions

English	Russian
- ***What*** *can I get you drink?* - A black coffee for me, please.	- ***И что*** *предложить Вам выпить?* - Чёрный кофе, пожалуйста.
- ***What*** *are you going to have to drink?* - I'd like something cool.	- ***А что*** *Вы желаете выпить?* - Хотелось бы чего-нибудь прохладного.
- ***What*** *are you going to have?* - A half of bitter, please.	- ***Что бы*** *Вы хотели сейчас?* - Полкружки горького, пожалуйста.
- ***What*** *is it to be?* - The same again, please.	- ***А что*** *теперь будете пить?* - То же самое, пожалуйста.

Figure 7 shows comparison of melodic portraits of special question intonations for English and Russian phrases. The figure layout and content representation is similar to **Figure 6**.



(a) English (b) Russian (c) English & Russian

Figure 7. UMPs for English and Russian one-accent-unit phrases (special questions)

The comparison of English and Russian special question melodic portraits allows to establish the following main differences:

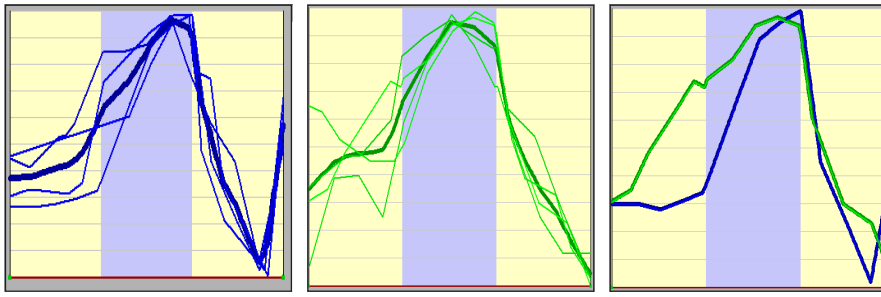
- the most significant changes are found in the pre-nucleus and nucleus regions;
- in the pre-nucleus region, the averaged Russian UMP is characterized by considerably higher level than the English UMP;
- in the nucleus region, the English UMP curve is characterized by a sharp rise in tonal frequency whereas the Russian curve remains on a steady high level;
- in the post-nucleus region, both Russian and English UMP curves demonstrate identical sharp interval decline.

Finally, for the study of intonation characteristics for general questions, we used the example English and Russian phrases listed in **Table 3**. The content representation and mark up in **Table 3** is similar to **Table 1**.

Table 3. English and Russian phrases spoken with the intonation of general questions

English	Russian
- *Does this **bus** go to the station?* - No, youll have to get off at the bank.	- *Этот автобус идет **на вокзал**?* - Нет, он идёт к банку.
- *Am **I** OK for St Marys Church?* - No, we only go as far as the park.	- *Я **правильно** иду к церкви?* - Нет, вы только дойдёте до парка
- ***Do** you go to the sea-front?* - No, youre going the wrong way.	- *Вы идёте к приморскому **бульвару**?* - Нет, Вы пошли неправильным путём.
- *Have we **got** much further to go?* - It's the next stop.	- ***Должны ли** мы ещё дальше ехать?* - Ваша остановка — следующая..

Figure 8 shows comparison of melodic portraits of general question intonations for English and Russian phrases. The figure layout and content representation is similar to **Figure 6**.



(a) English (b) Russian (c) English & Russian

Fig. 8. UMPs for English and Russian one-accent-unit phrases (general questions)

The comparison of English and Russian melodic portraits for general questions allows to establish the following main differences:

- the most significant changes are found in the pre- and post-nucleus regions;
- in the pre-nucleus region the English UMP curve is characterized by a steady low level, whereas the Russian UMP follows a substantial rise;
- in the post-nucleus region the English UMP curve is characterized by sharp decline with a subsequent rise towards the end of the phrase. On the other hand, the Russian UMP curve shows only steady decline;
- in the nucleus region, the English UMP curve shows a sharper rize in comparison to the English one.

Conclusions

The present paper describes the results of pilot studies where melodic portraits for English and Russian language phrases were compared. The study was restricted only to phrases with a one-AU structure representing the three main types of phrase intonations: *affirmative statements*, *special questions* and *general questions*.

The described results of comparisons of UMPs of English and Russian phrases are consistent with the observations of linguists involved in comparative studies of intonation in order to provide guidelines for mastering foreign languages. These guidelines often tend to have rather vague and descriptive language, for example:

"The melody of an English phrase differs markedly from a Russian one:

a). *The English voice range is much wider meaning that the beginning of the phrase is higher and the end of the phrase is lower in tone than in Russian.*

b). *English is characterized by the tonal movement within a vowel at a perceptibly longer time stretches which gives an impression of 'singing' stressed vowels.*

c). *The reference point of tone modulation in English is the lowest tone level while in Russian it is the average level.*

d). *The English cadence reaches the lowest point of the range, as well as tone rising from the lowest level.*

e). *The English phrase is characterized by the centralized accent. It is within the stressed syllable that the widest and longest voice cadence is exercised."*

(see: http://xreferat.com/71/1238-1-uprazhneniya-v-obuchenii-ritmu-i-intonacii-angliiyskogo-yazyka-v-osnovnoiy-shkole.html).

The described normalized UMP model of the phrase intonation allows to investigate the tonal differences between different languages by applying precise quantitative assessments. The method can be used effectively for solving problems of language interference. Moreover, the UMP model could potentially find an effective application in foreign language studies. Using the appropriate software that realizes the described stages of MP construction, a learner could be able to visually compare the intonation of the pronounced phrase with its target intonation portrait and work to eliminate a foreign accent by proper training.

The importance of mastering proper intonation in language instruction is emphasized by many authors:

"Intonation, the 'music' of a language, is perhaps the most important element of a correct accent. Many people think that pronunciation is what makes up an accent. It may be that pronunciation is very important for an understandable accent. But it is intonation that gives the final touch that makes an accent correct or native. Often we hear someone speaking with perfect grammar, and perfect formation of the sounds of English but with a little something that gives her away as not being a native speaker". (See http://www.goodaccent.com)

Another example. When taking about a Russian accent in American English some native speakers make interesting observations:

"Ask your average American what they think about the Russian accent and they say";

"Russians don't sound very friendly. I never feel as if they like me. I'm not sure if that's because of their language, or if it's a cultural thing".

One reason that Russian English speakers don't sound friendly is their flat tone. You simply don't use enough intonation when you speak.

Russian English speakers don't use the rising-falling intonation that Americans find friendly and engaging. You don't use sufficient intonation when asking questions".

(see: http://www.confidentvoice.com/blog/russian-english-speakers-5-reasons-why-americans-dont-understand-you/)

The author is grateful to **Dr. Anna Osipovich** for the useful discussions and for the help in preparation of English version of this paper.

References

1. *Lobanov B., Tsirulnik L., Zhadinets D., Karnevskaya E.* (2006) Language- and Speaker Specific Implementation of Intonation Contours in Multilingual TTS Synthesis // Speech Prosody: Proceedings of the 3rd International conference. Dresden, Germany: Vol. 2.—pp. 553–556.
2. *Lobanov B., Okrut T.* (2014) Universal Melodic Portraits of Intonation Patterns in Russian Speech [Universalnye melodicheskie portrety intonacionnyh konstrukciy russkoy rechi]// Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2014" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2006"], Bekasovo, pp. 330–339.
3. *Lobanov B. M.* (2015) An Experience of Melodic Portraits Creatiion of Complex Declarative Sentences of Russian [Opyt sozdaniya melodicheskih portretov povestvovatelnyh predlozheniy russkoy rechi] // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2015" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2015"], Moscow, pp. 414–426.
4. *Ockenden M ,* (2005) Situational Dialogues // The English Centre, Eastbourne / Revised Edition.—Longman—98 pp.
5. *R. Ogden, S. Hawkins, J. House, M. Huckvale, J. Local, P. Carter, J. Dankovicova, and S. Heid,* (2000) "Prosynth: an integrated prosodic approach to device-independent, natural-sounding speech synthesis," Computer Language and Science, pp. 177–210, 2000.

РАЗРЕШЕНИЕ ЛЕКСИЧЕСКОЙ МНОГОЗНАЧНОСТИ ДЛЯ РУССКИХ ГЛАГОЛОВ С ИСПОЛЬЗОВАНИЕМ СЕМАНТИЧЕСКИХ ВЕКТОРОВ И СЛОВАРНЫХ ДАННЫХ

Лопухин К. А. (kostia.lopuhin@gmail.com)
Scrapinghub, Москва, Россия

Лопухина А. А. (nastya-merk@yandex.ru)
Институт русского языка имени
В. В. Виноградова РАН, Москва, Россия.

Ключевые слова: полисемия, многозначность, разрешение лексической многозначности, частоты значений, семантические вектора

WORD SENSE DISAMBIGUATION FOR RUSSIAN VERBS USING SEMANTIC VECTORS AND DICTIONARY ENTRIES

Lopukhin K. A. (kostia.lopuhin@gmail.com)
Scrapinghub, Moscow, Russia

Lopukhina A. A. (nastya-merk@yandex.ru)
V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

Word sense disambiguation (WSD) methods are useful for many NLP tasks that require semantic interpretation of input. Furthermore, such methods can help estimate word sense frequencies in different corpora, which is important for lexicographic studies and language learning resources. Although previous research on Russian polysemous verbs disambiguation established some important and interesting results, it was mostly focused on reducing ambiguity or determining the most frequent sense, but not on evaluating WSD accuracy. To the best of our knowledge, there is no comprehensively evaluated method that can perform semi-supervised word sense disambiguation for Russian verbs. In this paper we present a WSD method for verbs that is able to reach an average disambiguation accuracy of 75% using only available linguistic resources: examples and

collocations from the Active Dictionary of Russian and large unlabeled corpora. We evaluate the method on contexts sampled from the web-based corpus RuTenTen11 for 10 verbs with 100 contexts for each verb. We compare different variations of the method and analyze its limitations. Method's implementation and labeled contexts are available online.

Key words: polysemy, word sense disambiguation, sense frequency, word2vec, semantic vectors

1. Introduction

Lexical-semantic ambiguity is an inherent property of any natural language, thus word sense disambiguation (WSD) in an important part of many natural language processing tasks. Various WSD techniques were discussed during SemEval sessions (Pradhan et al. 2007) and in WSD surveys (Ide and Véronis 1998; Navigli 2009; Mihalcea 2011). The most powerful and promising approaches are those that use already existing resources and do not require much human labeled data. Knowledge-based approaches take advantage of thesauri, for example English WordNet (Fellbaum 1998) or Russian Ru-Thes (Loukachevitch and Chujko 2007; Loukachevitch and Dobrov 2007) and encyclopedic resources, like Wikipedia (Ponzetto and Navigli 2010) and can be applied to domain-specific corpora with high accuracy (Agirre et al. 2009). Unsupervised corpus-based approaches typically perform clustering of senses in a corpus without making explicit references to any sense inventory, see e.g. (Schutze 1998; Huang 2012; Neelakantan 2014; Bartunov et al. 2015).

For Russian, several WSD experiments were performed on the Russian National corpus (RNC, ruscorpora.ru). Kobritsov et al. (2005) discussed the problem of automated word sense tagging in a large corpus and proposed a WSD approach based on lexical context markers. Shemanayeva et al. (2007) developed semantic filters aimed at raising the accuracy of sense disambiguation for adjectives in RNC. Mitrofanova et al. (2008) compared WSD techniques for Russian nouns that take into account either word lexical contexts or lexical-semantic tags of words in context. Both methods' average accuracies reached 83% and 85% respectively, which is comparable to the state-of-the-art in this field.

Sense disambiguation for Russian verbs was based on various linguistic resources. Kobritsov et al. (2007) performed a series of experiments on word sense disambiguation for verbs using information automatically extracted from dictionaries. They concluded that although the government pattern proves useful for disambiguation, automatically extracted information does not provide any substantial reduction of polysemy, and that accounting for semantic properties of the arguments is the most promising approach. Similar results were discussed in (Kustova and Toldova 2008). They studied several different methods of decreasing polysemy for Russian verbs: government pattern (morphological properties of arguments) and semantic properties of the arguments. Government pattern helped to halve the number of possible senses, and using fine-grained semantic properties of arguments allowed to reduce the number of possible senses to a single one for most studied verbs and most contexts.

Theoretical studies of verb polysemy prove that valencies and government patterns normally stay the same as regular metaphoric shifts take place (Rozina 2005; Reznikova 2014). Valencies usually change if new meanings appear in slang, for example *X gonit Y iz Z* 'X produces substance Y from substance Z' */ X gonit* 'X lies or says nonsense' from the first issue of the Active Dictionary of Russian (Apresjan et al. 2014). For our WSD experiments, which imply disambiguation for all verb senses, it seems reasonable to focus on lexical contexts of verbs rather than on their syntactic structure.

Other projects related to WSD of verbs are Disambiguation of Verbs by Collocation and Corpus Pattern Analysis led by Patrick Hanks and colleagues. These projects are focused on statistical analysis of corpus data in order to discover typical usage patterns and create the Pattern Dictionary of English Verbs (http://pdev.org.uk/; Hanks and Pustejovsky 2005; Hanks 2008). The authors emphasize that meanings are associated with prototypical sentence contexts (patterns or collocations) and not with word senses from dictionaries. Cf. also (Gries et al. 2010), where frequency distributions of English verbal constructions are discussed. The ongoing project of Russian FrameBank also focuses on verb constructions (http://framebank.ru/; Lyashevskaya 2012; Kashkin and Lyashevskaya 2013). Although the abovementioned methods presuppose disambiguation techniques and are corpus-based, they deal with collocations and not dictionary senses and we do not consider them in our study.

In this paper we set out to evaluate several techniques of word sense disambiguation for Russian verbs. The techniques are based on semantic vectors that use only existing linguistic resources—contexts and collocations from the Active Dictionary of Russian (AD; Apresjan et al. 2014). All the experiments are performed on 10 Russian verbs, whose senses are taken from AD. We evaluate all methods on contexts sampled from the web-based RuTenTen11 corpus (Kilgarriff et al. 2004). To the best of our knowledge, this is the first evaluation of a semi-supervised word sense disambiguation method for Russian verbs.

2. Method

The aim of our method is to be able to perform word sense disambiguation for Russian verbs using only existing linguistic resources, without any additional annotation. Such method needs a predefined sense inventory, and it is convenient if a single resource provides both senses and examples of their usage. In this paper we use sense inventory and examples from the Active Dictionary of Russian, a reliable resource with a strong theoretical basis in sense distinction that reflects contemporary language (Apresjan et al. 2014). Our disambiguation method consists of two major components: a context representation technique and a classifier trained on labeled contexts (examples and collocations from AD). More precisely, for each sense we extracted all examples (short and common usages), illustrations (longer, full-sentence examples from the Russian National Corpus), collocations, synonyms and analogues. Each example, illustration, etc. was treated as a separate context of a word used in a particular sense. Context representation technique takes contexts (some fixed

window of words before and after the disambiguated word) as an input and produces some real-valued vector as an output. This vector is then fed into the classifier, which predicts the sense of a context. Method implementation is available online[1].

There are a lot of options to choose from when building a context representation: whether to take word order into account, to parse the input sentence, to use lemmatization, to extract morphological features, how to represent words, etc. An important consideration is the amount of training data available, and the nature of the classification task. We have evaluated several options, but the most robust and performant for this task turned out to be representing context as a weighted average of individual word vectors. Word vectors are obtained by training a word2vec (Mikolov et al. 2013) model on a large lemmatized corpus (about 2 billion words—combined RuWac, lib.ru and Russian Wikipedia). Resulting vectors usually have 200–2000 dimensions and represent semantically similar words as vectors with similar directions. Using such vectors as input features allows us to leverage information from large unlabeled corpora and to generalize from one labeled context to all contexts that contain semantically similar words. We take a weighted average of individual word vectors: weights represent to which extent each word affects the sense of a context. Consider for example the word *verbovat'* (to recruit). If you see words such as *agent* 'agent' or *razvedka* 'intelligence service' in the context, these words alone give a strong hint about the sense of the target word. We give more weight to words that are more likely to be seen in the context of the target word than on their own (q_i here is the weight of the word). Negative weights are clipped to 0. If the word is unattested in available contexts, it is given a low weight of 0.2:

$$q_i = \begin{cases} \max\left(0, \ln \frac{P(w_i|c)}{P(w_i)}\right) & \text{if } P(w_i|c) > 0 \\ 0.2 & \text{if } P(w_i|c) = 0 \end{cases}$$

The context vector $\vec{c}$ is a weighted average of word in vectors, where words are taken from the window of 10 words before and after the target word, crossing sentence boundaries:

$$\vec{c_n} = \sum_{\substack{i=n-10 \\ i \neq n}}^{n+10} q_i \vec{w_i}$$

We have evaluated several classification approaches. In the first approach (Mean-Vec), we calculate an average of all context vectors for each sense during training, thus obtaining a single vector for each sense. When disambiguating an unlabeled context, we find a sense whose vector is the closest to the context vector. In the second approach (LR-Vec) we use an ensemble of logistic regression models with strong regularization. Each model is implemented as neural network that takes context vectors as input, applies dropout with p = 0.5 (Hinton et al. 2012), and a final softmax layer. An ensemble of 5 models is trained using stochastic gradient descent, and prediction of the most confident model is taken as the ensemble output. Dropout is employed to prevent overfitting on a small number of training contexts available from dictionary. The third approach (kNN-Vec) is a k-nearest neighbours classifier, where the

[1] https://github.com/lopuhin/sensefreq

sense is determined by k-closest neighbours of the input vector, and ties are broken using closeness between sense and context vectors. Best results were obtained with k = 3. Both Mean-Vec and kNN-Vec use cosine similarity as a closeness metric.

We also used two baseline methods (Bayes and SVM), as they are quite simple and often give strong results, especially with little training data. They used a different context representation: instead of semantic vectors, contexts were represented as bags of words with lemmatization and TF-IDF (term frequency inverse document frequency) downscaling. This means that inputs to these methods are sparse vectors, where each dimension corresponds to a word seen during training, and the only non-zero entries in each context vector correspond to words in this context vector. We compared two robust approaches for classification: a Bayes classifier and a support vector machine (SVM).

3. Evaluation

Word sense disambiguation accuracy is evaluated on contexts for 10 Russian verbs randomly sampled from the web-based RuTenTen11[2] corpus, the largest Russian corpus consisting of 18 billion tokens integrated into the Sketch Engine system (Kilgarriff et al. 2004). 7 verbs were randomly chosen polysemous verbs from the first issue of the Active Dictionary of Russian, and 3 (*vosprinimat'*, *brodit'*, *vypolzti*) were specifically chosen to make our test set of words maximally diverse. The word *brodit'* is homonymous (*Ona brodit po ulitse / Kvas brodit*) and both homonyms have several senses. Each word has 100 contexts labeled by one annotator. Annotated corpus is available online[3].

Comparison of word sense disambiguation accuracy on 10 verbs is presented in Table 1. For each word the table lists the number of senses, the ratio of the most frequent sense (MFS) in training data, and the disambiguation accuracy for 5 approaches. MFS is usually considered a strong baseline in supervised WSD (Navigli 2009), but this is not the case here: dictionary examples are used for training, and it is not known which sense is the most frequent. We see that approaches based on semantic vectors (the first three) perform significantly better, with LR-Vec giving the best average accuracy.

---

[2] Kutuzov and Kuzmenko (2015), Lopukhina, Lopukhin and Nosyrev (submitted) compared RNC and RuTenTen11 and found that they agree in most cases, including polysemous words.

[3] https://github.com/lopuhin/ruslang-wsd-labeled

Table 1: Word sense disambiguation accuracy on verbs from RuTenTen11 using Active Dictionary of Russian for training

Word	Senses	MFS	Mean-Vec	LR-Vec	kNN-Vec	Bayes	SVM
atakovat'	3	0.49	0.67	0.58	0.61	0.44	0.39
bayukat'	2	0.74	0.83	0.83	0.78	0.77	0.72
boltat'sya	5	0.37	0.55	0.60	0.47	0.45	0.28
bombardirovat'	4	0.47	0.83	0.83	0.85	0.64	0.55
brodit'	6	0.78	0.65	0.88	0.69	0.83	0.51
verbovat'	3	0.43	0.64	0.61	0.62	0.37	0.50
vlit'	3	0.89	0.91	0.90	0.92	0.64	0.65
volnovat'sya	3	0.96	0.97	0.97	0.97	0.96	0.94
vosprinimat'	3	0.70	0.77	0.78	0.78	0.71	0.65
vypolzti	6	0.42	0.54	0.48	0.42	0.36	0.30
Average	**3.8**	**0.624**	**0.735**	**0.745**	**0.712**	**0.617**	**0.548**

We also see that both baseline methods perform significantly worse than methods based on dense semantic vectors. This could be due to either the classification method or to the feature representation. In Table 2 we compare best performing semantic vector method (LR-Vec) with Bayes and SVM using two different feature representations (dense vectors and sparse vectors). Results suggest that the main gain is from the feature representation, and that LR method is able to use it more effectively.

Table 2: Comparison of average WSD accuracy across methods and feature representations

Method \ Features	Dense vectors	Sparse vectors
LR	**0.745**	0.600
Bayes	0.682	0.617
SVM	0.693	0.548

Some words in Table 1 have a much lower WSD accuracy: *atakovat'*, *boltat'sya*, *verbovat'* and *vypolzti*. Most of them (except for *atakovat'*) were difficult to the human annotator too, although we have no interrater agreement score to back this up. *Verbovat'* required very long contexts, often exceeding what was available, and general situation understanding. Disambiguation between several senses of *boltat'sya* required geometrical reasoning and sometimes quite specific knowledge. *Vypolzti* required animacy disambiguation, and often coreference resolution with long contexts. On the other hand, *brodit'*, which was by far the hardest word to annotate, has 88% accuracy with LR-Vec approach.

It is possible to analyse how our method determines the sense of the context. This is especially easy for the Mean-Vec approach: each sense has a vector, each word in a disambiguated context also has a vector and a weight, so we can calculate

closeness of a sense vector to each individual word vector. In Table 3 we present contexts for word *atakovat'* (to attack) with 3 senses: 'to attack someone in sport', 'to attack in a war conflict', and 'to ask someone many questions'. Word weight in the second column with values greater than 1.0 is highlighted in bold. The next three columns show similarities of individual words to each sense vector without taking weight into account. Context vector is calculated using individual word vectors and weights, and then compared with sense vectors, but since all operations are linear, it is equivalent to calculating weighted similarities for individual words, as shown in the table. It is interesting to observe that although the words *turki* (Turks) or *CSKA* (a sport team name) are absent in the training data, sense vectors help the method infer that the first word is more likely to be used in the war context, and the second in the sport context, due to their semantic similarity to other words in training data.

Table 3: Mean-Vec approach applied to an example context

	Weight	1: sport	2: war	3: questions
v	0.0	—	—	—
pervom	0.2	0.2	0.2	0.0
tajme	**3.0**	**0.4**	0.1	0.1
neskol'ko	0.5	0.1	0.1	0.0
opasnyh	**1.2**	**0.2**	0.1	0.0
momentov.	0.7	0.2	0.1	0.0
Prichem	0.2	0.2	0.1	0.0
turki	**2.3**	0.2	**0.3**	0.1
staralis'	0.2	0.1	0.0	0.0
bol'she	0.2	0.0	0.0	0.0
atakovat'	—	—	—	—
levym	**1.5**	**0.2**	**0.2**	0.0
flangom,	**3.9**	**0.5**	**0.4**	0.0
gde	0.0	—	—	—
iz-za	0.3	0.1	0.1	0.0
kadrovyh	0.0	—	—	—
problem	0.0	—	—	—
v	0.0	—	—	—
zashhite	**1.1**	0.1	**0.2**	0.0
u	0.0	—	—	—
CSKA	**1.6**	**0.4**	0.1	0.1
Result		**0.7**	0.4	0.1

Our method is targeted to be able to train on a relatively low number of diverse examples from the dictionary, but it is possible to evaluate it in a supervised manner. We can train it on 50 out of 100 labeled contexts for each word, using other 50 for testing. WSD accuracy in Table 4 is averaged on 10 random splits of data into training and test sets. All approaches perform better when trained on contexts from corpora

than on dictionary examples, but the gap is smaller than what we observed for nouns in (Lopukhina, Lopukhin and Nosyrev, submitted).

Table 4: WSD accuracy with different training data
(Active Dictionary of Russian and 50 contexts from RuTenTen11 corpus)

Training data	Mean-Vec	LR-Vec	kNN-Vec	Bayes	SVM
Active Dictionary	0.735	**0.745**	0.712	0.617	0.548
RuTenTen11	0.763	**0.778**	0.738	0.636	0.682

In the last part of this section we would like to analyze some implementation details and their effect on WSD accuracy (LR-Vec approach accuracy is reported unless otherwise specified):

- **Lemmatization**. It was not clear upfront whether lemmatization is required: on one hand, word2vec models achieve higher quality vectors from lemmatized corpora, on the other, lemmatization loses information that could be especially useful for verbs. It turns out that higher-quality word vectors are more important: WSD accuracy with lemmatization is significantly better: 0.738[4] vs. 0.709 with all other parameters being equal.
- **Stop-words**. Stop-words removal is a common pre-processing step in many NLP applications. But in this case stop-words can be important for disambiguation: for example, for *vosprinimat'* the WSD accuracy with stop-words removed is 0.68, whereas with stop-words included it is 0.78. Words such as *kak* turn out to be important discriminating factors here, while other stop-words are filtered out with weights. Stop-words were not removed when training a word2vec model, either.
- **Context size**. The context size determines which words will be included in the context vector used for classification. Our experiments show that almost all words benefit from larger contexts: average WSD accuracy is 0.745 with context size 10, 0.710 with 5 and 0.687 with 3, where context size 3 means that 3 words before and after the disambiguated word are used. An example of a larger context benefitting disambiguation is the context for *atakovat'* in Table 3: only words *tajm* and *CSKA* have high similarity with the correct sense, and they are far from the disambiguated word.
- **Accounting for word order**. All approaches described so far do not take word order into account, but word order is definitely useful for human speakers. We experimented with building contexts vectors by concatenating individual word vectors and using already described classifiers. But all methods we tried suffered from overfitting on larger context windows, with best results below 0.6 on window size 3. A possible way to resolve this problem is to build context representation that takes word order into account without supervision, and then use this representation to train a disambiguation method. We believe this is a promising direction for future study.

[4] This is different from 0.745 in Table 1 due to use of a smaller corpus and lower-dimensional vectors.

- **Using weights**. Using weight when building context vectors improves WSD accuracy, but less than what we observed for nouns in (Lopukhina, Lopukhin and Nosyrev, submitted): we achieve 0.741 without weights and 0.745 with weights.
- **Word2vec model**. We used the skip-gram model in all experiments. The main hyperparameters are window size and vector dimensionality. Larger window size (10–20) allows capturing topic/domain similarity, while smaller windows (2–3) better capture functional and syntactic similarity. In our experiments moderate window sizes (3–5) performed better: 0.723 with window size 10 and 0.745 with window size 5. Vector dimensionality also turned out to be an important factor: Mean-Vec and kNN-Vec significantly benefited from increasing it, while LR-Vec was less sensitive but still achieved best performance with 2048-sized vectors.

4. Conclusions

We presented a method that is able to reach a word sense disambiguation accuracy of 75% for Russian verbs. Our method uses context representation based on semantic vectors with weightening and dictionary information: examples and collocations from the Active Dictionary of Russian (Apresjan et al. 2014). Method's implementation and labeled contexts are available online on https://github.com/lopuhin/sensefreq.

Although our method shows good accuracy on many words, there are some words that are especially problematic. We believe that the general approach with semantic vectors is sound, but our current implementation is basic: it does not explicitly identify the verb's arguments, relying instead on weights to filter out words that help disambiguation. It would be interesting to check if a more sophisticated approach would improve the method's accuracy.

We would like to apply our method to other verbs and evaluate it on different types of predicates. The approach with the best performance will be implemented in the model for verb sense frequencies estimation, an ongoing project introduced in (Lopukhina, Lopukhin and Nosyrev, submitted)[5].

Acknowledgements

This work was supported by the grant from the Russian Scientific Foundation (No. 16-18-02054). We thank the anonymous reviewers for the many helpful comments and suggestions they made.

[5] http://sensefreq.ruslang.ru

References

1. *Agirre et al. 2009* — Agirre E. et al. Knowledge-Based WSD and Specific Domains: Performing Better than Generic Supervised WSD //IJCAI. — 2009. — Pp. 1501–1506.
2. *Apresjan et al. 2014* — Apresjan, Jury D., editor. Active Dictionary of Russian. A-G., Moscow. 2014.
3. *Bartunov et al. 2015* — Bartunov, Sergey, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. Breaking sticks and ambiguities with adaptive skip-gram. arXiv preprint arXiv:1502.07257. 2015.
4. *Fellbaum 1998* — Fellbaum, Christiane, editor. WordNet, An Electronic Lexical Database. The MIT Press, Cambridge, MA. 1998.
5. *Gries et al. 2010* — Gries S. T., Hampe B. and Schönefeld D. (2010). Converging evidence II: More on the association of verbs and constructions. In:Empirical and experimental methods in cognitive/functional research, CSLI Publications, pages 59–72.
6. *Hanks 2008* — Hanks P. Mapping meaning onto use: a Pattern Dictionary of English Verbs. In Proceedings of the AACL, Utah. 2008.
7. *Hanks and Pustejovsky 2005* — Hanks P. and Pustejovsky J. (2005). A Pattern Dictionary for Natural Language Processing. Revue Française de linguistique appliquée, 10(2):63–82.
8. *Hinton et al. 2012* — Hinton G. E. et al. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580. 2012.
9. *Huang 2012* — Huang E. H. et al. Improving word representations via global context and multiple word prototypes //Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. — Association for Computational Linguistics, 2012. — Pp. 873–882.
10. *Ide and Véronis 1998* — Ide N., Véronis J. Introduction to the special issue on word sense disambiguation: the state of the art // Computational linguistics. — 1998. — T. 24. — № 1. — Pp. 2–40.
11. *Kashkin and Lyashevskaya 2013* — Kashkin E. V., Lyashevskaya O. N. Semantic roles and construction net in russian FrameBank. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2013" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2013"], Bekasovo, pp. 827–836.
12. *Kilgarriff et al. 2004* — Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell. The Sketch Engine. Information Technology Research Institute Technical Report Series: pp. 105–116. 2004.
13. *Kobritsov et al. 2005* — Kobritsov B. P., Lyashevskaya O. N., Shemanayeva O. N. Lexico-semantic disambiguation in news and newspaper texts: surface filters and statistical estimation. [Snyatie leksiko-semanticheskoj omonimii v novostnyx i gazetno-zhurnal›nyx tekstax: poverxnostnye fil›try i statisticheskaya ocenka]. Internet-mathematics 2005: automated web-data processing. M.: 2005. Pp. 38–57.
14. *Kobritsov et al. 2007* — Kobritsov B. P., Lyashevskaya O. N., Toldova S. Ju. Semantic disambiguation of verbs using government patterns extracted from explanatory dictionaries. [Snyatie semanticheskoj mnogoznachnosti glagolov

s ispol›zovaniem modelej upravleniya, izvlechennyx iz e›lektronnyx tolkovyx slovarej]. Internet-mathematics — 2007, electronic publication, http://download.yandex. ru/IMAT2007/kobricov. pdf. — 2007.

15. *Kustova and Toldova 2008* — Kustova G. I., Toldova S. Ju. RNC: semantic filters for verb sense disambiguation. [NKRYa: semanticheskie fil'try dlya razresheniya mnogoznachnosti glagolov]. Russian National Corpus. — 2006. — T. 2008. — Pp. 258–276.
16. *Kustova G., Toldova S.* RNC: Semantic filters for the verb disambiguation ['NKRJA: semanticheskije filtry dlja razreshenija mnogoznachnosti glagolov']. Russian national corpus: 2006–2008. New results and perspectives. ['Natsionalnyi korpus russkogo jazyka: 2006–2008. Novye rezultaty i perspektivy']. Saint Petersburg: Nestor-Istorija. 2009.
17. *Kutuzov and Kuzmenko 2015* — Kutuzov, Andrey, and Elizaveta Kuzmenko. Comparing Neural Lexical Models of a Classic National Corpus and a Web Corpus: The Case for Russian. Computational Linguistics and Intelligent Text Processing. Springer International Publishing: pp. 47–58. 2015.
18. *Lopukhina, Lopukhin and Nosyrev, submitted* — Lopukhina A., Lopukhin K. and Nosyrev G. Automated word sense frequency estimation for Russian nouns. Submitted to Quantitative Approaches to the Russian Language (available online: http://sensefreq.ruslang.ru/download/Automated_Word_Sense_Frequency_Estimation_for_Russian_Nouns_Lopukhina_et_al.pdf)
19. *Loukachevitch and Chujko 2007* — Loukachevitch N. V., Chujko D. S. Automated lexical disambiguation using thesauri. [Avtomaticheskoe razreshenie leksicheskoj mnogoznachnosti na baze tezaurusnyx znanij]. Internet-mathematics 2007. Ekaterinburg: 2007. Pp. 108–117.
20. *Loukachevitch and Dobrov 2007* — Loukachevitch N. V., Dobrov B. V. Lexical disambiguation based on domain-specific thesaurus. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2007" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2007"], Bekasovo, pp. 400–407.
21. *Lyashevskaya 2012* — Lyashevskaya O. N. Dictionary of Valencies Meets Corpus Annotation: A Case of Russian FrameBank // Proceedings of EURALEX 15 (7–11 August 2012, Oslo, Norway). Oslo : Oslo University, 2012.
22. *Mihalcea 2011* — Mihalcea R. Word sense disambiguation // Encyclopedia of Machine Learning. — Springer US, 2011. — Pp. 1027–1030.
23. *Mikolov et al. 2013* — Mikolov T., Chen K., Corrado G., and Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013.
24. *Mitrofanova et al. 2008* — Mitrofanova O. A., Panicheva P. V., Lyashevskaya O. N. Statistical word sense disambiguation in contexts for names of physical objects. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2008" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2008"], Bekasovo, pp. 358–376.
25. *Navigli 2009* — Navigli, Roberto. Word sense disambiguation: A survey. ACM Computing Surveys (CSUR) 41.2: 10. 2009.

26. *Neelakantan 2014* — Neelakantan, Arvind, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014.
27. *Ponzetto and Navigli 2010* — Ponzetto S. P., Navigli R. Knowledge-rich word sense disambiguation rivaling supervised systems //Proceedings of the 48th annual meeting of the association for computational linguistics. — Association for Computational Linguistics. — 2010, pp. 1522–1531.
28. *Pradhan et al. 2007* — Pradhan, Sameer S., Edward Loper, Dmitriy Dligach, and Martha Palmer. SemEval-2007 task 17: English lexical sample, SRL and all words. Proceedings of the 4th International Workshop on Semantic Evaluations. Association for Computational Linguistics. 2007.
29. *Reznikova 2014* — Reznikova T. I. Author's style is decent, it really delivers: nonstandart semantic shifts in verbs (a study on internet texts). [Slog u avtora neplox, real›no dostavlyaet: o nestandartnyx semanticheskix perexodax v glagol›noj leksike (po dannym interneta)]. Modern Russian in the internet. M: 2014, pp. 169–181.
30. *Rozina 2005* — Rozina R. I. Semantic development in Russian literature and modern slang. [Semanticheskoe razvitie slova v russkom literaturnom yazyke i sovremennom slenge]. M.: Azbukovnik, 2005.
31. *Schutze 1998* — Schütze, Hinrich. 1998. Automatic word sense discrimination. Computational Linguistics 24, 1: pp. 97–124.
32. *Shemanayeva et al. 2007* — Shemanayeva O. Ju., Kustova G. I., Lyashevskaya O. N., Rakhilina E. V. Semantic filters for the word sense disambiguation in RNC: adjectives. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2007" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2007"], Bekasovo, pp. 582–588.

CREATING RUSSIAN WORDNET BY CONVERSION

Loukachevitch N. V. (louk_nat@mail.ru)[1],
Lashevich G. (design.berg@mail.com)[2],
Gerasimova A. A. (anastasiagerasimova432@gmail.com)[1],
Ivanov V. V. (nomemm@gmail.com)[2],
Dobrov B. V. (dobrov_bv@mail.ru)[1]

[1]Lomonosov Moscow State University, Moscow, Russia
[2]Kazan Federal University, Kazan, Russia

In this paper we have described the semi-automatic process of transforming the Russian language thesaurus RuThes (in version, RuThes-lite 2.0) to WordNet-like thesaurus, called RuWordNet. In this procedure we attempted to achieve two main characteristic features of wordnet-like resources: division of data into part-of-speech-oriented structures with cross-references between them and providing a set of relations similar to WordNet-like resources. The published version of RuWordNet contains more than 115 thousand Russian words and phrases presented in form of three lexical nets for nouns, verbs and adjectives. Between synsets such relations as hyponym-hypernym, meronymy, part-of-speech synonymy, antonymy are established. In the paper we compare web-page representations of RuThes 2.0 and RuWordNet. It can be seen that RuThes looks as an ontology describing concepts and their relations and RuWordNet looks as a net of words. Researchers can obtain both types of thesauri and compare them in applications. In future, we will continue to add new types of relations to RuWordNet including the domain relation, the cause relation, the entailment relation, etc.

Keywords: thesaurus, WordNet, natural language processing, lexical relations

СОЗДАНИЕ РУССКОГО WORDNET НА ОСНОВЕ КОНВЕРТАЦИИ ДАННЫХ ТЕЗАУРУСА РУТЕЗ

Лукашевич Н. В. (louk_nat@mail.ru)[1],
Лашевич Г. (design.berg@mail.com)[2],
Герасимова А. А. (anastasiagerasimova432@gmail.com)[1],
Иванов В. В. (nomemm@gmail.com)[2],
Добров Б. В. (dobrov_bv@mail.ru)[1]

[1]МГУ им. М. В. Ломоносова, Москва, Россия
[2]Казанский федеральный университет, Казань, Россия

Ключевые слова: тезаурус, WordNet, автоматическая обработка текстов, лексические отношения

1. Introduction

WordNet-like resources (Fellbaum, 1998) are one of the most popular resources used for natural language processing, wordnet projects have been initiated for many languages in many countries.

At least four attempts to create a Russian wordnet are known. RussNet (Azarowa, 2008) began development from scratch and at this moment appears to be quite small (not more than 20,000 synsets). Two other Russian wordnets were generated using automated translation (Gelfenbeyn et al., 2003; Balkova et al., 2008). The first one is publicly available[1] but represents the direct translation from Princeton WordNet without any manual revision. The last Russian wordnet project YARN (Yet Another Russian wordNet) was initiated in 2012 and is being created using a crowdsourcing approach; it currently contains mainly synsets with small number of relations between them (Braslavski et al., 2014).

For Russian, there exists the RuThes thesaurus, a linguistic ontology, which structure has differences from the WordNet approach. RuThes is a more ontology-oriented resource: thesaurus concepts have unique names, text entries of all parts of speech can be linked to the same concept, The RuThes relations are more formal conceptual relations. The current size of the published version of RuThes (RuThes-lite 2.0), accessible for non-commercial use, is more than 115 thousand text entries[2]. RuThes was specially created for information retrieval and natural language applications, it can be used in most applications where WordNet is usually utilized, but researchers and practitioners want to have a Russian wordnet.

In this paper, we describe the transformation of RuThes data to WordNet-like resource, called RuWordNet. In this process we try to reproduce two main features of the Princeton WordNet structure such as the organization in the form of part-of-speech lexical nets and the basic set of relations. The current volume of RuWordNet is the same as the published version of RuThes-lite 2.0 (115 thousand entries). It can be seen in Internet and can be obtained in the XML format.

The paper is organized as follows. The second section reviews the related work. The third section considers main features of the WordNet structure. The fourth section describes the main structure of RuThes and its differences from WordNet. The fifth section presents the transformation process from RuThes to RuWordNet and achieved results. The sixth section compares web-representations of RuThes and RuWordNet.

2. Related work

The most straightforward approach to the development of WordNet-like resources from scratch is a difficult task, which usually takes years of work. The approach to fasten the creation of a national wordnet is to translate Princeton WordNet

---

[1] http://wordnet.ru/

[2] http://www.labinform.ru/pub/ruthes/index.htm

to the target language (Vossen, 1998). Wordnet-like resources obtained with automatic translation can be generated fast enough but also requires a lot of efforts to be manually revised.

An intermediate approach between the above-mentioned ultimate points, which can be considered as quite usual, is to translate the top 5,000 concepts of the Princeton WordNet (core WordNet) and then extend this hierarchy manually, using local dictionaries. This approach was accepted in the development of EuroWordnet (Vossen, 1998) and Danish wordnet—DanNet (Pedersen, 2010).

Analysing previous approaches for national wordnet development, authors of FinnishWordNet (FiWN) decided to use manual translation of Princeton WordNet synsets by professional translators. The direct translation approach was based on the assumption that most synsets in PWN represent language-independent real-world concepts. Thus, the semantic relations between synsets were also assumed mostly language-independent, so the structure of PWN could be reused as well. In such a way, Finnish wordnet, FinnWordNet (FiWN), was created by translating more than 200,000 word senses in the English Princeton WordNet (PWN) 3.0 in 100 days.

Braslavski et al (2014) suppose to create a Russian wordnet (YARN) utilizing Russian Wiktionary and crowdsourcing.

Wiktionary is a crowdsourced dictionary and thesaurus that exists for many languages. Wiktionary pages related to a specific word can contain a lot of useful information about word senses, including a list of lexical senses, definition and examples for a lexical sense, lexical relations (synonyms, antonyms, hyponyms, hypernyms), which are represented as links to Wiktionary pages. However, there are also some problems in word senses description, which can hamper creating a WordNet-like resource especially for inexperienced crowdsourcers:

- a lexical link leads not to a specific sense but to the whole word page ,
- synonyms can be described as partial synonyms, this is a very vague notion: *гейзер, фонтан* [*gayser, fountain*].
- lexical relations are not symmetrical. For example, word w_1 is indicated as a synonym to word w_2, but word w_2, is not indicated as a synonym to word w_1. In other examples, word w_1 is indicated as a synonym to word w_2, but word w_2 is indicated as a hypernym to word w_1.

3. Basic Structure of Princeton WordNet

The structure of Princeton University's WordNet (and other wordnets) is based on sets of partial synonyms—synsets, organized in hierarchical part-of-speech-based lexical nets for nouns, adjectives, verbs, and adverbs. Each part-of-speech net has its own system of relations between synsets.

The most frequent relation between noun synsets is the hyponym-hypernym relation. Also since 2006 in Princeton WordNet class-instance relations denoted as Instance Hypernym and Instance Hyponym (Miller, Hristea, 2006) were introduced. Such relations substituted hyponym-hypernym relations for synsets of proper nouns

designating unique entities such as cities, countries, concrete persons, etc. This work was made under the influence of the ontologists' point of view on "confusion between individuals and concepts" (Gangemi et al., 2001).

The noun relationships also include part-whole relations, which are subdivided into proper part-whole relations (*wing* is a part of *bird*), member parts (*tree* is a member of *forest*), and material (*snow* is a substance of *snowball*). Parts can have several wholes (*wing* is a part of *bird*, *bat*, *insect*, or *angel*).

For all parts of speech, the lexical relation of antonymy can be established. Lexical relations link lexemes, not whole synsets.

In Princeton Wordnet, the antonymy relation is the main type of relations for descriptive adjectives (Gross, Miller, 1990), which were described only with the relations of antonymy and similarity. For example, for the word *heavy*, the word *light* is indicated as an antonym, such words as *hefty, ponderous, massive* are linked to *heavy* with the relation "similar to". Other wordnets, such as GermaNet (Kunze, Lemnitzer, 2010) or Polish WordNet (PlWordNet) (Derwojedowa et al., 2008), changed this approach and introduced taxonomic relations (hyponymy-hyperonymy) for adjectives.

Verbs in WordNet are mainly linked with hyponym-hypernym relations. Besides, they have their own unique relations not described for nouns or adjectives: entailment (*buy—pay*) and causation (*give—have, kill—die*). The WordNet entailment relation is a relation between two verbs V_1 and V_2 that holds when the sentence "*Someone* V_1" logically entails "*Someone* V_2" and there is the temporal inclusion of event V_1 into V_2 or vice versa (Fellbaum, 1998). The causation relation can be also considered as a subtype of a general logical entailment relation but there is not temporal inclusion between corresponding situations (Fellbaum, 1998).

4. RuThes Structure and Relations

RuThes and WordNet are both thesauri that are lexical resources where semantically related words and expressions are collected together into synsets or concepts between which formalized relations are set. When applying both thesauri to natural language processing, the same steps should be made such as matching between a text and a thesaurus and employing the described thesaurus relations if necessary. The most evident differences between the two types of resources are as follows.

First, in RuThes there is no division into subnets according to different parts of speech that is words of any part of speech can be linked to the same concept if they mean the same (so called derivative or part-of speech synonyms).

Therefore, second, in RuThes it is often very difficult or even impossible to apply traditional tests of synonymy detection such as substitution of synonyms in sentences (Cruse, 1986, Miller, 1998). Tests checking the denotational scope of lexemes are usually applied in the following way: "if entity X can be called with word W_1, then we can call it also with word W_2" and vice versa regardless of specific context. The second test consists in formulation of explicit differences of one concept from other

concepts. These differences can be fixed in the unique concept name. Thus, abovementioned issues of RuThes such as denotational tests, denotational distinctions between concepts, and unique names of concepts make RuThes much closer to ontological resources in an imaginary scale from lexical resources to formal ontologies than WordNet-like thesauri. RuThes can be called a linguistic (lexical) ontology for natural language processing.

Third, the relations in RuThes are only conceptual, not lexical (as antonyms or derivational links in wordnets). They are constructed as more formal, ontological relations of traditional information-retrieval thesauri (Z39.19, 2005), which were designed to describe very broad, unstructured domains. The set of conceptual relations includes:

- the class-subclass relation;
- the part-whole relation applied with the following restriction: the existence of the concept-part should be strictly attached to the concept-whole. For example, trees can grow in many places not only in forests therefore concept *TREE* cannot be directly linked to concept *FOREST* with the part-whole relation, the additional concept *FOREST TREE* should be introduced;
- the external ontological dependence when the existence of a concept depends on the existence of another concept (in such a way forests depend on the existence of trees) (Guarino, Welty, 2002). In RuThes we denote this relation as association with indexes: asc_1 is directed to the main concept, asc_2—to the dependent concept;
- In the very restricted number of cases symmetric associations between concepts can be established.

The main idea behind this set of relations is to describe the most essential, reliable relations of concepts, which are relevant to various contexts of concept mentioning. Also this set of relations allows us to describe domain terminologies or domain-specific ontologies, combine descriptions of lexical and domain-specific knowledge in the same resource.

The relation of ontological dependence is very convenient for describing conceptual relations between concepts corresponding to multiword expressions and concepts of their component words (such as *nature protection* and *nature*), which allows easier introducing such concepts and describing useful "horizontal" relations.

Thus, RuThes has considerable similarities with WordNet: the inclusion of concepts based on senses of real text units, representation of lexical senses, detailed coverage of word senses. At the same time the differences include attachment of different parts of speech to the same concepts, formulating of names of concepts, attention to multiword expressions, the set of conceptual relations, etc. The more detailed description of RuThes and RuThes-based applications can be found in (Loukachevitch, Dobrov, 2014) or (Lukashevich, 2011).

At present RuThes includes 54 thousand concepts, 158 thousand unique text entries (75 thousand single words), 178 thousand concept-text entry relations, more than 215 thousand conceptual relations. The published version of RuThes, RuThes-lite 2.0, contains 115 thousand text entries. It was singled out from full RuThes on the basis of words and phrases used in current Russian news flows with exclusion several specific domains (Loukachevitch et al., 2014).

5. Conversion from RuThes to RuWordNet

According to the guidelines of world-known WordNet thesaurus, the first version of Russian wordnet (RuWordNet) was created.

In our opinion, one of the most distinctive features of WordNet-like resources is their division into synset nets according to parts of speech. Therefore all text entries of RuThes-lite 2.0 were subdivided into three parts of speech: nouns (single nouns, noun groups, or preposition groups), verbs (single verbs and verb groups), adjectives (single adjectives and adjective groups). We have obtained 29,297 noun synsets, 12,865 adjective synsets, and 7,636 verb synsets (Table 1).

This subdivision was based on the morphosyntactic representation of RuThes-lite 2.0 text entries, which was fulfilled semi-automatically. Therefore, a small number of mistakes because of particle treatment (verbs or adjectives) or substantivated adjectives can appear. For example, Russian phrase *любитель подраться* (=*драчун*) [*brawler, scrapper*] was treated in this procedure as a verb group and currently is assigned to the verb synsets. Currently all found mistakes are corrected.

The divided synsets were linked with the relation of part-of-speech synonymy.

Table 1. Quantitative characteristics of synsets in RuWordNet

Part of speech	Number of synsets	Number of unique entries	Number of senses
Noun	29,296	68,695	77,153
Verb	7,634	26,356	35,067
Adjective	12,864	15,191	18,195

The hyponym-hypernym relations were established between synsets of the same part of speech. These relations include direct hyponym-hypernym relations from RuThes-lite 2.0. In addition, the transitivity property of hyponym-hypernym relations was employed in cases when a specific synset did not contain a specific part of speech but its parent and child had text entries of this part of speech. In such cases the hypernymy-hyponymy relation was established between the child and the parent of this synset.

Similar to the current version of Princeton WordNet, in RuWordNet class-instance relations are also established. By now, they had been generated semi-automatically for geographical objects.

The part-whole relations from RuThes were semi-automatically transferred and corrected according to traditions of WordNet-like resources. Now RuWordNet contains 3.5 thousand part-whole relations. The part-whole relations include the following subtypes:

- functional parts (*nostrils—nose*),
- ingredients (*additives—substance*),
- geographic parts (*Sevilia—Andalusia*),
- members (*monk—monastery*),
- dwellers (*Moscow citizen—Moscow*),
- temporal parts (*gambit—chess party*)
- inclusion of processed, activities (*industrial production—industrial cycle*)

Adjectives in RuWordNet similarly to German or Polish wordnets are connected with hyponym-hypernym relations. For example, word *цветовой* [*colored*] is linked to such hyponyms as *красный* [*red*], *синий* [*blue*], *зеленый* [*green*], etc.

Adjectives often have POS-synonymy links to nouns, but also can have POS-synonyms to verb synsets. For example, word *строительный* has two POS-synonymy relations: to the noun synset {*стройка, постройка, возведение, сооружение*..} and to the verb synset {*строить, построить, возводить* ...}.

The specific feature of the current state of adjectives description in RuWordNet is the existence of part-whole relations (*портовый—прибрежный*) and even instance-class relations (*майкопский—северо-кавказский*) (see Table 2), which adjectives inherited from RuThes concepts. These relations should be renamed to hyponym-hypernym relations.

Table 2. Number of different relations in RuWordNet

Part of speech	Hypernyms	Instance-Class	Wholes	POS-synonymy	Antonyms
Noun	39,155	1,863	10,010	18,179	455
Verb	10,440	0	117	7,451	20
adjective	17,834	66	829	14,139	457

In the current RuWordNet representation of Russian verbs, part-whole relations can be seen. For example, synset {*видеть во сне, сниться, грезиться, присниться, привидеться во сне, пригрезиться, пригрезиться во сне*»} [*to dream*] is linked to synset {*спать, поспать, доспать, соснуть, досыпать, почивать, проспать, просыпать*} [*to sleep*] with the part-whole relation. Such a relation between the translation equivalents [*to dream—to sleep*] exists also in Princeton WordNet and called 'entailment relation'. Another example from RuWordNet is {*оппонировать, оппонировать диссертацию*}, which is described as a part for {*защитить диссертацию*}. Christian Fellbaum wrote in (Fellbaum, 1998) that «the entailment relation between verbs resembles meronymy between nouns, but meronymy is better suited to nouns than to verbs». Thus, the simple renaming of the part-whole relations between verbs in RuWordNet into entailment relations is possible and correct.

Antonymy relations are conceptual relations in RuWordNet, that means they link synsets, not single lexemes. They are introduced for all parts of speech, mainly for synsets denoting properties and states, for example:

- noun synset {*легкость, с легкостью, без труда, без затруднений*} [*ease* as noun] is antonymous to synset {*тяжесть, трудность*} [*difficulty*],
- adjective synset {*легкий, легкий для выполнения, легкий для осуществления, нетрудный*} [*ease* as adjective] is antonymous to synset {*тяжкий, трудный, тяжелый, трудный для выполнения, нелегкий ...* } [*difficult*],
- verb synset {*не соответствовать действительности*} [*to be contrary to the fact*] is antonymous to synset {*соответствовать истине, соответствовать действительности*} [*to be in accordance with the truth*].

The current numbers of relations described in RuWordNet are presented in Table 2.

6. Publication of RuThes and RuWordNet on the Web

RuThes-lite 2.0 and RuWordNet are published in form of static web-pages. Looking through RuThes[3], the user should select a letter to begin, next select an initial trigram of a word, and then click on a proper word. For example, selecting word *двор* [*yard*] the user can find three concepts associated with this word, relations of these concepts, and other text entries attached to the same concepts. Further, the navigation through concepts or text entries is possible (Fig. 1).

In the similar representation of RuWordNet[4], there is the initial division to parts of speech, which the user should select, then the user should find a word. In the RuWordNet representation, there are no concepts (Fig. 2), each synset contains text entries belonging to the same part of speech, POS-synonymy links to other parts of speech are indicated. Thus, in the representation RuThes looks more as an ontology, and RuWordNet is presented more as a lexical net.

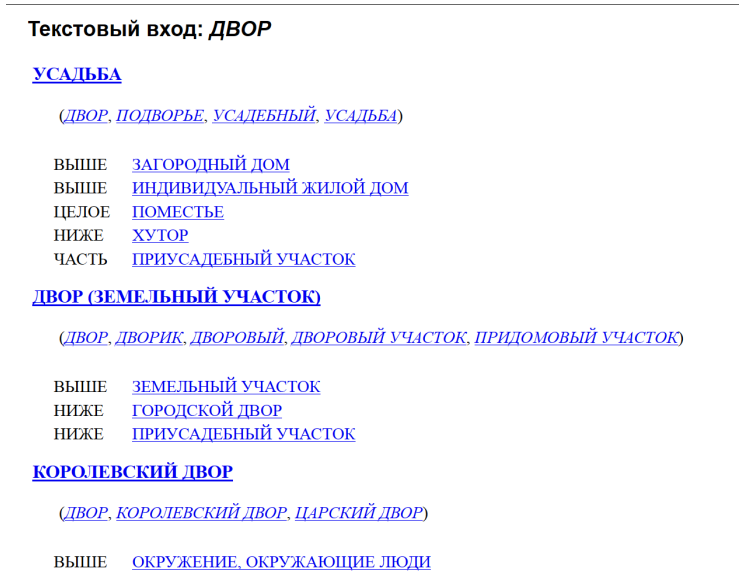


Fig. 1 Representation of three senses of the Russian word *двор* in RuThes

[3] http://www.labinform.ru/pub/ruthes/index.htm

[4] http://www.labinform.ru/pub/ruwordnet/index.htm

Fig. 2. Representation of senses of Russian noun *двор* in RuWordNet: synsets contain only nouns, concept name are not presented, there are references to POS synonyms (adjectives)

Conclusion

In this paper we have described the semi-automatic process of transforming the Russian language thesaurus RuThes (in version, RuThes-lite 2.0) to WordNet-like thesaurus, called RuWordNet. In this procedure we attempted to achieve two main characteristic features of wordnet-like resources: division of data into part-of-speech-oriented structures with cross-references between them and providing a set of relations similar to wordnet-like relations.

Both thesauri, RuThes-lite 2.0 and RuWordNet, are currently published as static web-pages. Also RuWordNet can be seen through web interface[5]. Researchers can obtain both types of thesauri, compare them in applications. In future, we will continue to add new types of relations to RuWordNet including the domain relation, the cause relation, the entailment relation, etc.

[5] http://ruwordnet.ru

Acknowledgments

This work is partially supported by Russian Foundation for Humanities grant № 15-04-12017.

References

1. *Azarowa I.* (2008), RussNet as a Computer Lexicon for Russian, Proceedings of the Intelligent Information systems IIS-2008, pp. 341–350.
2. *Balkova V., Suhonogov A., Yablonsky S.* (2008), Some Issues in the Construction of a Russian WordNet Grid, Proceedings of the Forth International WordNet Conference, Szeged, Hungary, pp. 44–55.
3. *Braslavski P., Ustalov D., Mukhin M.* (2014), A Spinning Wheel for Yarn: User Interface for a Crowdsourced Thesaurus, In Proceedings of EACL-2014, Gothenberg, Sweden, pp. 101–104.
4. *Cruse D.* (1986), Lexical Semantics, Cambridge, University Press.
5. *Derwojedowa M., Piasecki M., Szpakowicz S., Zawisławska M., Broda B.* (2008), Words, concepts and relations in the construction of Polish WordNet, In Proceedings of the Global WordNet Conference, Seged, Hungary, pp. 162–177.
6. *Fellbaum C.* (1998), A semantic network of English verbs, WordNet: An electronic lexical database, pp. 153–178.
7. *Gangemi A., Guarino N., Masolo C., Oltramari A.* (2001), Understanding Top-Level Ontological Distinctions, Proceedings of IJCAI 2001 Workshop on Ontologies and Information Sharing, pp. 26–33.
8. *Gelfenbeyn I., Goncharuk A., Lehelt V., Lipatov A., Shilo V.* (2003), Automatic translation of WordNet semantic network to Russian language, Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog-2003.
9. *Gross D., Miller K. J.* (1990), Adjectives in WordNet, International Journal of Lexicography, 3(4), pp. 265–277.
10. *Guarino N., Welty C.* (2002), Evaluating ontological decisions with ONTOCLEAN, Communications of the ACM, 45(2), pp. 61–65.
11. *Kunze C., Lemnitzer L.* (2010), Lexical-Semantic and Conceptual relations in GermaNet, In Storjohann P (ed) Lexical-semantic relations: Theoretical and practical perspectives, pp. 163–183.
12. *Lindén K., Niemi J.* (2014), Is it possible to create a very large wordnet in 100 days? An evaluation, Language resources and evaluation, 48.2, pp. 191–201.
13. *Loukachevitch N., Dobrov B.* (2014), RuThes Linguistic Ontology vs. Russian Wordnets. In Proceedings of the Seventh Global WordNet Conference (GWC 2014), pp. 154–162.
14. *Loukachevitch N. V., Dobrov B. V., Chetviorkin I. I.* (2014), RuThes-Lite, a publicly available version of thesaurus of Russian language RuThes, In proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog-2014, pp. 340–350.

15. *Lukashevich N. V.* (2001), Thesauri in information-retrieval tasks, Moscow (in Russian).
16. *Miller G.* (1998), Nouns in WordNet, In WordNet: An Electronic Lexical Database, Fellbaum, C (ed), The MIT Press, pp. 23–47.
17. *Miller G. A., Hristea F.* (2006), WordNet nouns: Classes and instances, Computational linguistics, 32(1), pp. 1–3.
18. *Pedersen B. S., Nimb S., Asmussen J., Sørensen N. H., Trap-Jensen L., d Lorentzen, H.* (2009), DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary, Language resources and evaluation, 43(3), pp. 269–299.
19. *Vossen P.* (1998), Introduction to EuroWordNet. In EuroWordNet: A multilingual database with lexical semantic networks, Springer Netherlands, pp. 1–17.
20. *Z39.19*. (2005), Guidelines for the Construction, Format and Management of Monolingual Thesauri, NISO.

SENTIRUEVAL-2016: ПРЕОДОЛЕНИЕ ВРЕМЕННЫХ РАЗЛИЧИЙ И РАЗРЕЖЕННОСТИ ДАННЫХ ДЛЯ ЗАДАЧИ АНАЛИЗА РЕПУТАЦИИ ПО СООБЩЕНИЯМ ТВИТТЕРА

Лукашевич Н. В. (louk_nat@mail.ru)
МГУ им. М. В. Ломоносова, Москва, Россия

Рубцова Ю. В. (yu.rubtsova@gmail.com)
Институт систем информатики им. А. П. Ершова СО РАН, Новосибирск, Россия

Ключевые слова: анализ тональности текстов, классификация текстов по тональности, социальные сети, разметка коллекций

SENTIRUEVAL-2016: OVERCOMING TIME GAP AND DATA SPARSITY IN TWEET SENTIMENT ANALYSIS

Loukachevitch N. V. (louk_nat@mail.ru)
Lomonosov Moscow State University, Moscow, Russia

Rubtsova Y. V. (yu.rubtsova@gmail.com)
A. P. Ershov Institute of Informatics Systems, Novosibirsk, Russia

In this paper we present the Russian sentiment analysis evaluation SentiRuEval-2016 devoted to reputation monitoring of banks and telecom companies in Twitter. We describe the task, data, the procedure of data preparation, and participants' results. At the previous evaluation SentiRuEval-2015, it was noticed that the presented machine-learning approaches significantly depended on the training collection, which was not enough for qualitative classification of the test collection because of data sparsity and time gap. The current results of the participants at SentiRuEval-2016 showed that they have made successful steps to overcome the above-mentioned problems by combining machine-learning approaches and additional manual and automatically generated lexical resources.

Keywords: sentiment analysis, sentiment classification, social network, collection labeling, evaluation

1. Introduction

One of the important directions in automatic sentiment analysis is the analysis of social network messages, especially Twitter posts (tweets). Twitter messages convey a lot of opinions on various topics written by people of different origin, education, employment, etc, which can be interesting to governments, sociologists, companies, and ordinary people.

Twitter messages have several specific features. They are short (140 symbols), and their content is dynamic, often very dependent on current events. For this reason, automatic sentiment classifiers, trained in a restricted set of tweets, significantly lose in their quality if applied to tweet collections of other time intervals.

In [1, 2] the analysis of participants' results in the Russian tweet task of SentiRuEval-2015 was presented. Comparing results in two subtasks: sentiment analysis (reputation monitoring) towards telecommunication companies and towards banks, it was shown that best achieved levels of results significantly correlated with the differences between training and test collection. In that competition, the training and test collections were divided with the half-year interval, during which dramatic Ukraine events happened and partially changed the topics of the tweets. The analysis of the most problematic tweets for the participants showed that such tweets (30% in the bank domain) included sentiment words absent in the training set.

During this year, the second evaluation of tweet-oriented sentiment analysis systems was organized at SentiRuEval-2016. In this paper, we describe the task, the principles of data annotation, the achieved results and present the best approaches, which tried to overcome time-related problems of the tweet sentiment analysis.

2. Related Work

In past years several shared tasks were devoted to sentiment analysis and reputation monitoring of opinionated tweets.

In 2012–2013 RepLab, online reputation management evaluation, was held within the CLEF conference [3, 4]. The task was to determine if the tweet content has positive or negative implications for the company's reputation. The RepLab organizers emphasize that the RepLab task is substantially different from standard sentiment analysis that should differentiate subjective from objective information. When analyzing polarity for reputation, both facts and opinions have to be considered to determine what implications a piece of information might have on the reputation of a given entity. The training and test collections were temporally divided with at least several month intervals.

To overcome the difference between the training and test collections, the participants combined supervised approaches with unsupervised approaches or lexicon-based approaches. Some runs incorporated external information by using provided links to Wikipedia, entities' official web sites, and external vocabularies.

The highest F-measure and accuracy values among RepLab 2013 were achieved by the system SZTE NLP [5]. The team utilized the external vocabularies: the

SentiWordNet sentiment lexicon [6] and the acronym lexicon (from www.internetslang.com). Beyond the supervised steps, they experimented with unsupervised clustering using Latent Dirichlet Allocation (LDA) for detecting topics in the training and test collections. Then they used the topic distributions over each tweet as features.

The second best system according F-measure and third one according Accuracy was POPSTAR [7]. The team used sentiment lexicons to extract features based on the prior polarity of words. Some tweet-oriented features were included to capture particular aspects of tweets (e.g. presence of emoticons). The participant claims that delta-tf.idf weight scheme for word features shows the best results for this task. To solve the problem of feature vector sparseness and unseen words, they implemented the Brown cluster algorithm that clusters words to maximize the mutual information of bigrams.

The approach of the UAMCLYR [8] was based on distributional term representations (DTRs) [9], which are a way to represent terms by means of contextual information, given by term-co-occurrence statistics. The participant demonstrated that the proposed approach shows better result in comparison to the traditional Bag-of-Words representation.

In 2013–2015, the Twitter-oriented sentiment evaluation was held within the SemEval conference. Two subtasks were given to participants in 2013–2014: to detect sentiment expressed by a phrase in the context of a tweet and to detect overall sentiment of a tweet [10, 11]. In 2015 organizers included three new subtasks asking to predict the sentiment towards a topic in a single tweet, the overall sentiment towards a topic in a set of tweets, and the degree of prior polarity of a phrase [12].

In SemEval 2013 and 2014, the best result by a large margin was shown by NRC-Canada system [13]. The sentiment lexicon features (both manually created and automatically generated) along with n-gram features (both word and character n-grams) led to the most achievement in performance. Additionally, they generated two large sentiment association lexicons, one from tweets with sentiment-word hash tags, and another one from tweets with emoticons.

The best system for Subtask C (prediction sentiment towards a given topic) at SemEval 2015 was TwitterHawk [14]. The team focused on identifying and incorporating the strongest features used by the best systems in previous years, most notably, sentiment lexicons that showed good performance in earlier studies. Their system used two kinds of features: basic text features and lexicon features. They have incorporated eight external lexicons. To increase the classifier quality, they extended the training collection with the data from subtask B.

In 2015 the first SentiRuEval evaluation of Russian sentiment analysis systems was held [1, 2]. The aim of the tweet analysis was to classify messages according to their influence on the reputation of the mentioned company. The analysis of the participants' results showed that the best achieved performance in the reputation oriented-task for a specific domain was correlated with the difference between word probability distributions over the training and test collections in this domain. From the description of the approaches, it became clear that no additional data (word clusters or lexicons) were not used by the participants in their supervised machine-learning methods.

3. SentiRuEval-2016 Twitter Task

Similar to the previous SentiRuEval-2015 evaluation, the goal of the Twitter sentiment analysis at SentiRuEval-2016 was to find tweets influencing the reputation of a company in two domains: banks and telecom companies. Such tweets may contain sentiment-oriented opinions or positive and negative facts about the company.

Such a task is quite similar to the reputation polarity task at RepLab evaluation [3, 4] and sub-task C in SemEval 2015. The difference from RepLab evaluation is that at SentiRuEval, tweets from only two domains were taken, and the systems were evaluated for these domains separately, which gives the possibility to compare the results obtained in the domains. The task for participants was to define the reputation-oriented attitude of a tweet in relation to a given company: positive, negative, or neutral.

In the training and test collections, the fields with the list of all companies of the chosen domain were denoted. By default, the field of the company mentioned in the tweet obtained "0" (neutral attitude) value. The participants should either replace "0" with "1" (positive attitude), or "–1" (negative attitude), or remain "0", if the tweet attitude to a company mentioned in the message is neutral.

3.1. Text Collections

The SentiRuEval collections comprise tweets about seven entities from the telecom domain and eight entities from the bank domain. The datasets were collected with Streaming API Twitter (https://dev.twitter.com/streaming/overview). The previous SentiRueval-2015 training and test collections (December 2013—January 2014; July-August 2014) were utilized as training collections of the current evaluation. The current test collections were gathered in two parts: during July 2015 and November 2015. The distribution of messages in the training and test collections according to sentiment classes is shown in Table 1. The number of tweets is not equal to the sum of neutral, positive and negative messages, as user may mention more than one company in a message. As it can be observed, the collections are unbalanced and we did not artificially boost the number of sentiment tweets—just how classifiers would face the sentiment classification task in the real life.

Table 1. The distribution of messages in the collections according to polarity classes in the SentiRuEval datasets

		Neutral	Positive	Negative	Total number of tweets
Telecom	Training collection	4,870	1,354	2,550	**8,643**
	Gold standard test collection	1,016	226	1,054	**2,247**
Banks	Training collection	6,977	704	1,734	**9,392**
	Gold standard test collection	2,240	312	722	**3,313**

The Twitter task of SentiRuEval-2016 was to determine the reputation-oriented attitude of a tweet in relation to a given company. Some tweets could contain more than one entity. Table 2 displays the number of tweets that contain more than one company and the number of tweets with different polarity labels.

To prevent manual labeling by participants, additional messages have been added to the test collections. The size of the collections sent to the participants was equal to 19,673 tweets for the Telecom domain and 19,586 tweets for the Bank domain.

Table 2. Number of tweets that contain more than one company

		Number of tweets containing more than one entity	Number of tweets containing different polarity labels
Telecom	Training collection	435	131
	Gold standard test collection	193	49
Banks	Training collection	857	23
	Gold standard test collection	101	11

3.2. Data Annotation and Quality Measures

A high-quality gold standard collection is essential for supervised machine learning. Traditionally the gold standard is created by expert annotators. However, traditional annotation is expensive and time-consuming. To reduce the cost of expert-based annotation, linguistic projects have turned to the crowdsourcing approach, which involves submitting smaller subtasks to a coordinated platform on the Internet and solving these smaller tasks with a large amount of people. Nowadays crowdsourcing is becoming an increasingly popular and rather practical approach for creation and annotation of linguistic resources [15,16]. Crowdsourcing can employ both paid workers and volunteers.

In the framework of SentiRuEval-2016, the online tool (http://sentimeter.ru/assess/texts/) for tweet labeling was created where one could mark tweets according to their attitude in relation to a given company. 8,509 tweets in total were loaded into the system and labeled by assessors: 3,970 tweets about telecommunication companies and 4,539 tweets about banks. The labeling process lasted since 1 September 2015 to 31 January 2016. The interface of the online crowd source platform for sentiment labeling is shown on Fig. 1.

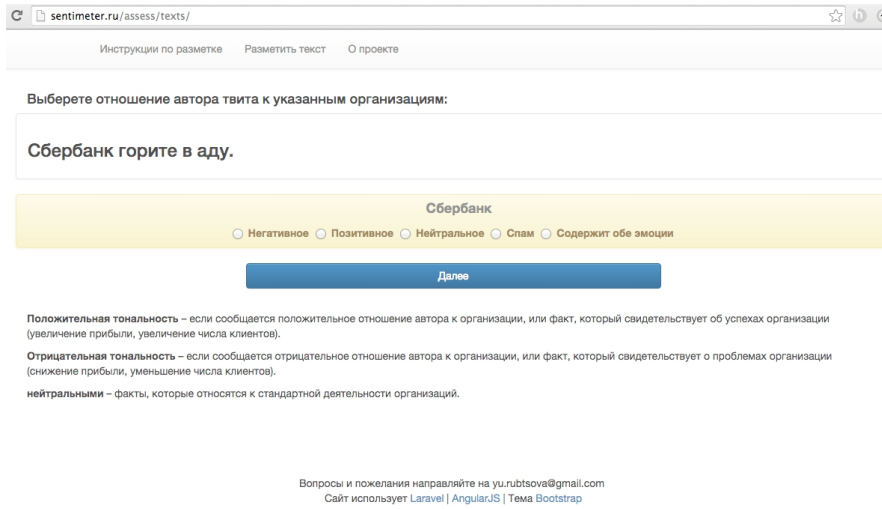


Figure 1. The interface of the online crowdsource platform for tweet sentiment labeling

For the four-month period of assessment, total 112 people from 25 cities and 7 countries took part in labeling. The annotators can be subdivided into three different groups: organizers—two persons, paid assessors—four persons, and volunteers—106 persons. All together they marked 45,450 companies. The organizers marked 10,322 companies, the paid assessors labeled 29,435 companies, and the volunteers labeled 5,693 companies (approximately 54 companies per volunteer).

To reduce the subjectivity, each tweet from the test collection was marked by at least four different persons as a person may feel preference or antipathy to some brand or company and mark tweets prejudiced. For instance, if a person is a "brand advocate" then he or she can label tweets as "neutral" if there is the slightest possibility not to label it as "negative".

After labeling was finished, the "strong agreement" voting scheme was applied to form the test collections. The labeling of a tweet was considered to be in strong agreement if the number of votes for a specific sentiment label exceeded votes for other labels with the margin 2. So, a tweet was filtered out from the gold standard if three assessors voted for one mark and two ones for another one—it was assumed as disagreement. Only tweets with strong agreement among assessors have formed the gold standard. Irrelevant, unclear, or spam messages were removed from the test sets.

As the main quality measure, macro-average F-measure was used. Macro F-measure is calculated as the average value between F-measure of the positive class and F-measure of the negative class ignoring the neutral class. But similar to SentiRuEval-2015, this does not reduce the task to the two-class prediction because erroneous labeling of neutral tweets negatively influences Fpos and Fneg. Additionally, micro-average F-measures were calculated for two sentiment classes.

4. Results and Description of Approaches

This year ten participants have submitted 58 runs to the Twitter sentiment analysis task at SentiRuEval-2016. The best runs according to macro-F for each participant are presented in Table 3 for telecom tweets and Table 4 for bank tweets.

In the evaluation we calculated two baselines. The first baseline is based on the major reputation-oriented category—negative one in both cases. The best runs of all participants show results above the F-macro majority baseline, however, some systems could not surpass the F-micro baseline.

The second baseline is obtained with the use of SVM to Boolean representation of tweet wordforms (if a wordform is presented in a tweet then the feature is equal to 1, otherwise 0). Six of ten participants could beat this baseline. If compared to SentiRuEval-2015, the considerable improvement can be seen because at the previous evaluation the best approaches in the bank domain were at the level of the SVM baseline (F-macro=0.3578, F-micro=0.3736). In the telecom domain, the best results were better than the SVM baseline, but the current margin between the SVM baseline and the best result is bigger (baseline: F-macro=0.4396, F-micro=0.48; the best result: F-macro=0.488, F-micro=0.536).

Two most popular machine-learning approaches among participants were SVM and neural networks. To overcome the differences between the training and test collections, five best approaches used machine learning in conjunction with external resources. Two participants (1 and 10) tried to increase the classification results by balancing the train collections. Three participants (1, 9, and 10) incorporated external sentiment vocabularies into supervised machine learning algorithms.

Table 3. The best run from each participant for telecom tweets according Macro F

	F-macro	F-micro
Majority Baseline	0.3146	0.5895
SVM baseline	0.4640	0.5728
1_4	0.5286	0.6632
2_k	**0.5594**	0.6569
3_1	0.3634	0.3994
4_5	0.4955	0.6252
5_1	0.3499	0.4044
6_con	0.3545	0.5263
7_5_a	0.4842	0.6374
8_533_2	0.4871	0.5745
9_hand_ext_tri	0.5493	**0.6813**
10_10	0.5055	0.6254

Table 4. The best run from each participant for banks tweets according Macro F

	F-macro	F-micro
Majority Baseline	0.1885	0.3503
SVM-baseline	0.4555	0.4952
1_4	0.4683	0.5022
2_k	**0.5517**	**0.5881**
3_1	0.3423	0.3524
4_1	0.376	0.4108
5_1	0.3859	0.464
6_con	0.2398	0.3127
7_5_a	0.471	0.5128
8_533_2	0.4492	0.4705
9_auto_ext_tri	0.5245	0.5653
10_5	0.4659	0.5053

These sentiment lexicons were as follows: the manual lexicon SentiRuLex[1] [17], the automatically generated lexicon study.mokoron.com [18], and the crowdsourced lexicon—Linis crowd[2] [19]. Participant 2 generated word clusters on a large collection of social network posts and comments and utilized them in tweet classification.

Below we briefly describe the best (compared to the baselines) approaches employed by the SentiRuEval participants. Participant 1 used words in uppercase, bigrams and punctuation marks as features for the linear-kernel SVM. The participant also integrated extra lexicons based on the following collections: study.mokoron.com [18], the collection of tweets (January 2016), and manual lexicon RuSentiLex [17]. The training collections were extended with other tweets in order to balance them. The telecom balanced collection consisted of 4,894 tweets, the banking balanced collection consisted of 6,980 tweets.

Participant 2 employed the recurrent neural network, and the long short-term memory (LSTM) model in particular. As features, Participant 2 used word2vec trained on the external collection of social network posts and comments.

The participant 9's best result for telecom companies is based on SVM over unigram, bigrams, and trigrams. Additionally, two vocabularies were implemented into the classifier: RuSentiLex [17] and automatic connotation vocabularies generated from a news collection. The best approach of this participant for the bank tweets also was based on SVM with the same features as it was used for the telecom domain but only the connotation lexicon was used and the consideration of the part-of-speech ambiguity was added.

The best runs of Participant 10 for the telecom tweets and banks also differ, but the only difference is that the classifier for telecom tweets worked better with a stop-word list, which showed the poor results for the bank domain. The participant used linear SVM with tweet-specific normalizations and integrated the RuSentiLex lexicon. The tweet-specific normalizations mean that the participle "not" plus a word was considered as one feature; multiple characters were replaced by a two-fold repetition; links, replies, dates, numbers were replaced with patterns.

The distribution of results of all 58 runs for the telecom test collection can be observed on graph 1. Graph 2 shows the distribution among all runs for the bank test collection.

We analyzed tweets that were incorrectly classified by all participants and found that most such tweets mention several entities with different sentiments, for example:

> *"А я вам всегда говорил, что лучший сотовый оператор это Билайн. Мегафон вас не уважает".* [I always said to you that the best operator is Beeline. Megaphone does not respect you].

Another found problem concerns phrase sentiment. The following tweet (and its several variants) was considered by all systems as negative:

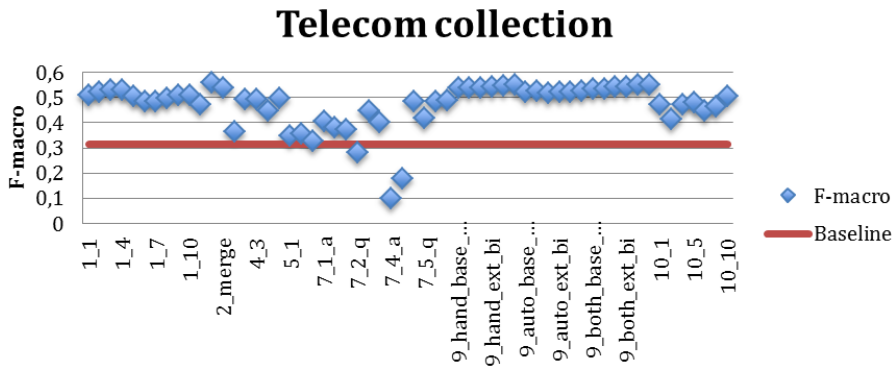
> *"ВТБ 24* ***сократил убыток*** *вдвое во II квартале"* [VTB-24 **reduced losses** in II quarter].

---

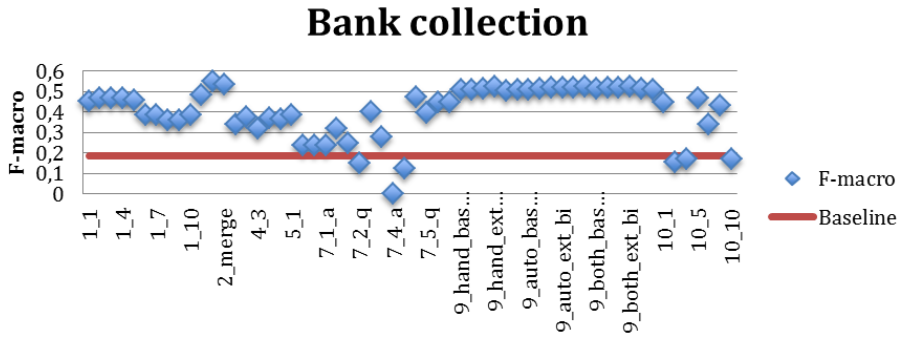
[1] http://www.labinform.ru/pub/rusentilex/index.htm

[2] http://linis-crowd.org/

Thus, it seems that dependence of the best systems from a training collection decreased, the systems now can use a lot of additional information. But they also should extract additional knowledge about phrase sentiment and try to find better ways to analyze different attitudes in the same tweet.



Graph 1. The distribution of all runs for Telecom collection



Graph 2. The distribution of all runs for Bank collection

Conclusion

In this paper we presented the Russian sentiment analysis evaluation SentiRuEval-2016 devoted to reputation monitoring of banks and telecom companies in Twitter. We described the task, data, the procedure of data preparation, and participants' results. At the previous evaluation SentiRuEval-2015, it was noticed that the presented machine-learning approaches significantly depended on the training collection, which was not enough for qualitative classification of the test collection because of data sparsity and time gap. The current results of the participants at SentiRuEval-2016 showed that they have made successful steps to overcome the

above-mentioned problems by combining machine-learning approaches and additional manual and automatic lexical resources.

All prepared collections are available for research purpose https://goo.gl/GhX3vU.

Acknowledgments

This work is partially supported by RFBR grants No. 14-07-00682 and No. 15-07-09306.

References

1. *Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Y., Ivanov V., Tutubalina E.* (2015), SentiRuEval: testing object-oriented sentiment analysis systems in Russian. Proceedings of International Conference Dialog-2015, pp. 3–9.
2. *Loukachevitch N., Rubtsova Y. Entity-Oriented Sentiment Analysis of Tweets: Results and Problems* (2015), Proceedings of Text-Speech-Dialog-2015, LNAI, Springer, 9302, pp. 551–559.
3. *Amigo E., Corujo A., Gonzalo J., Meij E., de Rijke M.* (2012), Overview of RepLab 2012: Evaluating Online Reputation Management Systems, CLEF 2012 Evaluation Labs and Workshop Notebook Papers, Rome.
4. *Amigo E., Albornoz J. C., Chugur I., Corujo A., Gonzalo J., Martin T., Meij E., de Rijke M., Spina D.* (2013), Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems, CLEF 2013, Lecture Notes in Computer Science Volume 8138, pp. 333–352.
5. *Hangya V., Farkas R.* (2013), Filtering and Polarity Detection for Reputation Management on Tweets, CLEF-2013 Working Notes.
6. *Baccianella, S., Esuli, A., Sebastiani, F.* (2010), SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D., eds.: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).
7. *Filgueiras J., Amir S.* (2013), POPSTAR at RepLab 2013: Polarity for Reputation Classification, CLEF-2013 Working Notes.
8. *Villatoro-Tello E., Rodríguez-Lucatero C., Sánchez-Sánchez C., López-Monroy A. P.* (2013), UAMCLyR at RepLab 2013: Profiling Task. In CLEF (Working Notes).
9. *Lavelli A., Sebastiani F., Zanoli, R.* (2004), Distributional Term Representations: An Experimental Comparison. In Italian Workshop on Advanced Database Systems.
10. *Nakov P., Kozareva Z., Ritter A., Rosenthal S., Stoyanov V., Wilson T.* (2013), Semeval-2013 task 2: Sentiment analysis in Twitter, Proceedings of the 7th International Workshop on Semantic Evaluation SemEval-2014.
11. *Rosenthal S., Ritter A., Nakov P., Stoyanov V.* (2014), SemEval-2014 Task 9: Sentiment Analysis in Twitter, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, pp. 73–80.

12. *Rosenthal S., Nakov P., Kiritchenko S., Mohammad S., Ritter A., Stoyanov V.* (2015), Semeval-2015 task 10: Sentiment analysis in twitter. Proceedings of SemEval-2015. Denver, Colorado, June 4–5, 2015. Association for Computational Linguistics, pp. 451–463.
13. *Mohammad S., Kiritchenko S., Zhu X.* (2013), NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In Proceedings of the Second Joint Conference on Lexical and Computational Semantics (SEMSTAR'13).
14. *Boag W., Potash P., Rumshisky A.* (2015), TwitterHawk: A Feature Bucket Approach to Sentiment Analysis. SemEval-2015, pp. 640–646.
15. *Bocharov V., Alexeeva S., Granovsky D., Protopopova E., Stepanova M., Surikov A.* (2013), Crowdsourcing morphological annotation. In Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue", RGGU, pp. 109–124.
16. *Braslavski P., Ustalov D., Mukhin M.* (2014), A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus, Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics.—Gothenburg, Sweden : Association for Computational Linguistics, pp. 101–104.
17. *Loukachevitch N., Levchik A.* (2016), Creating a General Russian Sentiment Lexicon. In Proceeding of LREC-2016.
18. *Rubtsova Y.* (2015), Constructing a corpus for sentiment classification training, "Programmnye produkty i sistemy" (Software & Systems), №1 (109), pp. 72–78.
19. *Alexeeva, S., Koltsov, S., Koltsova O.* (2015), Linis-crowd.org: A lexical resource for Russian sentiment analysis of social media, Computational linguistics and computational ontology, pp 25–34.

LEXICAL RESEARCH IN RUSSIAN: ARE MODERN CORPORA FLEXIBLE ENOUGH?

Lukashevich N. Y. (natalukashevich@mail.ru)
Moscow State University, Moscow, Russia

Klyshinsky E. S. (klyshinsky@mail.ru)
Keldysh IAM RAS, Moscow, Russia

Kobozeva I. M. (kobozeva@list.ru)
Moscow State University, Moscow, Russia

The article discusses what modern tools offer for a corpus-based lexical research in Russian. As an example we analyzed how the adjective *gordy* 'proud' is used in modern news texts. We studied data from such resources as two general Russian language corpora (RNC, GICR) and a corpus of syntactic co-occurrences containing information on syntactic relations of words for Russian (CoSyCo[1]). If a corpus includes a variety of genres and allows to make fine-grained distinctions between text sources, it helps to highlight important style- and genre-dependent differences. Our comparison has demonstrated that there are quite significant differences in the usage of *gordy* which become clear when we study general news and IT news corpora separately, however, in general they show certain similar tendencies. It is also shown that when more varied genres are taken into account it may make more visible such style and genre features which it is not so easy to notice otherwise.

Key words: corpus-based research, flexibility, words co-occurrence, Russian lexis, lexical semantics

НАСКОЛЬКО ГИБКИ КОРПУСА ДЛЯ ЦЕЛЕЙ ИССЛЕДОВАНИЯ РУССКОЙ ЛЕКСИКИ

Лукашевич Н. Ю. (natalukashevich@mail.ru)
МГУ им. М. В. Ломоносова, Москва, Россия

Клышинский Э. С. (klyshinsky@mail.ru)
ИПМ им. М. В. Келдыша РАН, Москва, Россия

Кобозева И. М. (kobozeva@list.ru)
МГУ им. М. В. Ломоносова, Москва, Россия

[1] This project is partially funded by RFH grant 15-04-12019.

В работе обсуждаются возможности, предлагаемые современными средствами для корпусных исследований лексики русского языка. В качестве примера анализируется употребление прилагательного *гордый* в современных новостных текстах на материале данных двух общих корпусов русского языка (НКРЯ, ГИКРЯ) и корпуса синтаксической сочетаемости, содержащего информацию о синтаксических связях слов в русском языке (КОСИКО). Сравнение показывает, что, несмотря на наличие общих тенденций, есть значительные различия в употреблении слова *гордый* в текстах компьютерных новостей и новостей общей тематики. Делается вывод, что наличие текстов разных жанров и возможность проводить разграничения между ними с точностью до источника позволяет увидеть существенные жанровые и стилистические различия, не столь заметные при рассмотрении материала в общем.

Ключевые слова: корпусные исследования, гибкость, совместная встречаемость слов, русская лексика, лексическая семантика

Much has been said in the ongoing discussion what kind of corpus a linguist/lexicographer needs. Such features of a corpus as its representativeness, volume, accessibility, etc. have been widely discussed. Recently the concept of register variation (Belikov et al, 2013) came into focus. It has been mentioned that this approach which allows to account for heterogeneity of language data is relevant not only for sociolinguistic studies, but also for a whole lot of linguistic tasks including comparative studies of texts from different genres (Belikov et al, 2012).

The issue of how to apply this approach to the sphere of genre and style is not that simple. For different research we may need different degrees of what might be called genre and style granularity in a corpus. Besides, a researcher may not be aware in advance of where and what kind of differences the study will reveal, so ideally (s)he needs a possibility to tune this granularity in accordance with the current demands.

In this paper we show that the ability to make fine-grained distinctions between sources in a corpus and to compare texts of similar but not identical genres may be of crucial importance.

1. Existing corpora and flexibility

It is a fact that modern linguistic and lexicographic research is mostly conducted on corpus data. Such studies depend much of the quality of corpora, the type(s) of phenomena covered by these resources, their structure, their limitations, etc.

It is also generally understood that the more a corpus allows tailoring the initial data and search in accordance with the needs of a particular study, the more varied tasks can be solved with its help.

Sites of modern corpora offer a variety of tools aimed to help researchers. Quite a lot of resources allow to refine selection features of words in the search string. These include Russian National Corpus (RNC) (Lashevskaja, Plungian, 2003), the General Internet Corpus of Russian (GICR) (Belikov et al, 2012), Sketch Engine (Kilgariff et al., 2004), etc.

Sketch Engine provides word sketches—corpus-based summaries of the grammatical and collocational behaviour of a selected word.

RNC Sketches[2] project also generates sketches using syntactically tagged texts from RNC; its resulting output contains information on syntactical relations between words.

Until quite recently the existing corpora for Russian did not offer a way to create a subcorpus of your choice. However, at the moment several corpora provide this opportunity to a researcher.

One of them is GICR, which was developed as a resource that allows to apply the method of segmental statistics in a wide range of linguistic tasks (Belikov et al, 2012). Its interface makes it possible to select not only the corpus segment, but also the type of text sources according to the list of segment-specific attributes (such as the author's year of birth, place of birth, gender for blogs, the name of the source for news, etc).

Another one, RNC has previously only allowed to select text sources from the main subcorpus on the basis of certain features. However, quite recently a similar possibility appeared in other subcorpora including its newspaper subcorpus[3].

In this paper we will focus on how using this option affects the results of a lexical study.

2. Data collection

Our aim was to study the usage of the word *gordy* 'proud' in such sphere as news texts. We have analysed how the word is used in the first 1,000 contexts of the main subcorpus of RNC, its subcorpus of newspaper texts, and the news segment in GICR.

RNC's newspaper subcorpus (as of 12/12/2015) included texts of the following 7 sources:

Total	**173.5 mln (words)**	**100%**
Izvestia	9,282,250	5.35%
Komsomolskaya Pravda	44,867,100	25.86%
Novy region	25,174,850	14.51%
RBK Daily	26,424,050	15.23%
RIA Novosti	15,545,600	8.96%
Sovetsky sport	12,977,800	7.48%
Trud	39,228,350	22.61%

[2] (http://ling.go.mail.ru/synt/)

[3] Unfortunately, this happened after the data for this research was collected and analysed, so we were not able to include it in this paper.

GICR news segment (as of 03/02/2016) included texts from the four sources as follows:

Source	**851 mln (words)**	**100%**
lenta	110,418,290	12.97%
regnum	218,025,910	25.60%
ria	337,669,976	39.65%
rosbalt	185,456,412	21.78%

For both corpora we obtained contexts with *gordy* followed directly by an (in) animate noun. For GICR we used search queries with the switched on by default option of deleting duplicates in the results.

Besides these two, we used a self-designed collection of texts, which is flexible to the highest degree as a result. We took as such a collection of the news subcorpus of CoSyCo—a corpus of syntactic co-occurrences containing information on syntactic relations of words for Russian, which is currently being developed for the purposes of teaching Russian as a foreign language.

CoSyCo news subcorpus contains texts from the following sites:

News sites:	**982.2 mln (words)**	**100%**
lenta.ru (lenta)	71,300,115	7.26
RBK (rbc)	61,933,721	6.31
RIA Novosti (ria)	409,971,920	41.74
Nezavisimaya gazeta (ng)	48,923,879	4.98
Vzglyad (vz)	72,370,767	7.37
Rossiyskaya gazeta(rg)	71,467,194	7.28
Commersant (commersant)	140,585,843	14.31
Polit.ru (polit)	49,697,364	5.06
Utro.ru(utro)	45,770,623	4.66
Ibusiness.ru (ibusiness)	10,131,894	1.03
IT news	**113 mln (words)**	**100%**
Мембрана (membrana)	7,391,018	6.40
CNews (cnews)	35,830,813	31.04
Компьютерра (computerra)	28,068,619	24.32
Компьюлента (compulenta)	16,204,248	14.04
PCWeek (pcweek)	27,924,230	24.19

Since our newspaper collection was not tagged, we used the software tool we created during the development of CoSyCo database[4]. The site cosyco.ru provides an easy access to a database of such syntactically connected word combinations as adjective+noun and verb+preposition+noun. The paper (Klyshinsky et al., 2011) describes the method that was used to create this database. In the current project, we have created a tagger for a more flexible clause extraction. The developed software tool processes more than 15 mln word tokens per minute, thus it is very convenient

[4] http://cosyco.ru/

for small collections (e.g. texts of one newspaper for one year). Processing large corpora (e.g. the whole Librusec collection[5]) takes about 12–16 hours. The tool works as a Windows command-line local application taking input in an XML-file, which contains a query and a list of input and output files.

Following such tools as NLTK and Stanford Parser, our software allows writing regular-expression-like queries including an ambiguity analysis. Words in input queries can have not just one but several initial forms or parts of speech. Like the Universal Dependencies[6] initiative, we separate feature name and its value. Such separation of name and value of features helps to make the denotation of words coordination or its absence relatively simpler. For example, the query

```
(in;prep;)(;adv;)(;adj;gender+, number=pl, case+ & ;noun;
 case-)(;adj;case- & ;noun;gender+,number=pl, case+)
```

matches a clause that includes the preposition *'in'*, an adverb followed by two words that are ambiguous for part of speech (between adjective and noun). Two last words are in plural form **(number=pl)**; the first adjective and second noun are coordinated by gender and case **(gender+, case+)**, while the first noun and the second adjective **(case-)** are not coordinated. The main features of the developed query language are described in (Vlasova et al, 2016), however, in this project we used a slightly different notation.

We ran our tool over the selected collection using a simple query

```
(ГОРДЫЙ;adj; case+, gender+, number+)(;adj; case+, gender+,
number+)*(;noun; case+, gender+, number+).
```
[7]

This means the word *gordy* in all its forms that is possibly followed by an iteration of adjectives; the clause is finished by a noun; all words have the same values of case, gender, and number. Our tests showed that this query has high recall and precision. We selected the first 1,000 of sentences from the RNC's main subcorpus output for the word gordy. Our tool returned 364 sentences with just 8 mistakes found by the assessor and 100% recall. Mistake rate on our newspaper collection varies from 1% up to 13% with the average at 5%.

Out of the relevant contexts for each resource we compiled a list of nouns, which co-occurred with *gordy* in its texts.

These lists of nouns were analyzed from the point of view of semantic classes which could be identified there. At the same time the correlations between the semantic class of a noun and the semantic interpretation (sense) of its adjectival modifier *gordy* were studied.

---

[5] http://lib.rus.ec/

[6] http://universaldependencies.org/

[7] As for duplicates (which are a problem for some of the sources used), they had to be deleted manually, as currently there is no automatic deduplication in CoSyCo. (By duplicates here we mean exact copies of sentences, as there were also cases when only a part of the sentence was repeated (mostly when official political comments were reported)—such cases were counted as separate instances.)

3. Meanings of gordy and semantic classes of co-occurring nouns

During the analysis of 1,000 contexts from RNC's main subcorpus we came to distinguish 5 senses of *gordy*[8]:

1) *gordy X* ≈ a person X whose behavior shows that (s)he has a sense of dignity and self-respect:

   (1) *(Katya) suddenly felt doubly happy: her beloved was not an ordinary man, no, he was tough, proud and pure.*[9] [RNC, E. Kazakevich, Zvezda][10]

2) *gordy X* ≈ a person X who is feeling pleased with the fact that smth that (s)he (or someone associated with him/her) owns or smth (s)he (or the associates) achieved should make other people think better of him/her or rank him/her higher:

   (2) *Alevtina is proud that she earns her living herself and does not depend on anybody…*[11] [RNC, V. Makanin, Otdushina]*[*

3) *gordy X* ≈ a person X who thinks of oneself as being better than other people and treats them with contempt because of that:

   (3) *You had better go and stay with the guests, or they will think you are too proud*[12]. [RNC, A. Chekhov, V rodnom uglu]

4) (figurative) majestic, stately

   (4) *High, proud celestial mountain peaks were glimmering golden in the sunset sky.*[13] [RNC, V. Skripkin, Tinga]

5) (figurative) sublime, lofty, elevated:

---

[8] These senses correspond to slightly modified three senses of this word given in MAS. Unlike MAS, which unites all figurative meanings under *gordy2* sense, we actually singled them out as separate ones (namely, *gordy* 4 and 5) and added them to the three MAS senses related to person.

[9] (Катя) вдруг почувствовала себя вдвойне счастливой: ее любимый был не обычный человек, нет, он суровый, гордый и чистый. [Э. Г. Казакевич. Звезда (1946)]

[10] Authors' translation—here and below, except for (5).

[11] Алевтина горда тем, что зарабатывает на жизнь сама и ни от кого не зависит …. [В. Маканин, Отдушина (1977)]

[12] Ты бы посидела с гостями, а то подумают, что ты гордая. [А. П. Чехов. В родном углу (1897)]

[13] Высокие, гордые вершины небесных гор румянились в закатном небе. [В. Скрипкин. Тинга // «Октябрь», 2002]

(5) *Fyodor thought…with proud, joyous energy, with passionate impatience, he was already looking for the creation of something…* [RNC, V. Nabokov. The Gift (M. Scammel, V. Nabokov, 1962)][14]

The resulting list of semantic classes of nouns modified by *gordy* (with minimal examples) is presented below. It is divided into four groups on the basis of combinability with the 5 senses of the adjective.

1. Nouns denoting persons:
   - a person in general or a male/female person:
     gordaya devushka 'a proud girl'
   - a person according to family status
     gordy otets 'a proud father'
   - a person according to their nationality/ethnicity:
     gordy amerikanets 'a proud American'
   - a person according to their social status
     gordy korol' 'a proud king'
   - a big group of people (community)
     gordy narod 'a proud people'

etc.

2. Nouns related to the situation of a person being proud (in sense 2):
   - nouns denoting a person as a possessor of smth or an agent of some deed:
     gordy pobeditel' 'a proud winner'
   - nouns denoting emotions experienced by a person who feels proud (in the sense 2):
     gordaya radost' 'proud joy'

3. Nouns referring to an object which is similar to a proud person in some respect:
   - an inanimate object:
     gordaya bashnya 'a proud tower'
   - an animal, a bird:
     gordy oryol 'a proud eagle'

4. Nouns denoting features of a person or results of human activity:
   - general characteristics of a person (as a subject of mental and social activity):
     gordiy um 'a proud mind'
   - emotional states, feelings, personality traits:
     gordoe spokoistvie 'proud tranquility'
     gordoe prezrenie 'proud contempt'
   - features and qualities of appearance and movement:
     gordaya osanka 'proud demeanor'
   - psychological states, processes, and 'products' of mental work

---

[14] Федор Константинович …с какой-то радостной, гордой энергией… уже искал создания чего-то нового… [В. В. Набоков, Дар (1935–1937)]

gordaya mechta 'a proud dream'
- representational objects (linguistic units and expressions)
  gordoye imya 'a proud name'
- *gordaya nadpis*' 'a proud inscription'

etc.

Nouns in the first group above are supposedly used with the first or the third meaning of *gordy* (depending on whether the speaker assesses the given instance of behaviour positively or negatively), and the second group combines with the second meaning of *gordy*[15].

The third group of nouns covers all cases of metaphoric transfer when an object is compared to a proud person (realization of *gordy4*); the comparison is often based on the look typical for a proud person (i.e. holding one's head high, not bending, etc).

The fourth group includes cases (linked with *gordy5*) when the characteristic is transferred by metonymy from a proud person (1, 2 or 3) to something which can express pride (as a personality trait or emotion). Here we find not only "inherent" features of a person (related to appearance, mind, character, etc), but also "results" of social, mental, emotional activity of a person.

4. Results

To evaluate the variance for the word *gordy*, we calculated a feature vector including all nouns which co-occurred with *gordy*. In corpus-based research, the feature vector usually contains instances per million (ipm) value (Lyashevskaya and Sharoff, 2009). However, here we are interested in the changes in the frequency distribution. That is why we took the feature vector containing absolute values of words co-occurrences and normalized it on the sum of frequencies, i.e. calculated the conditional probability of meeting a noun in a context with the word *gordy*.

We started from the idea that by normalizing data across different sources (or groups of sources) we lose information on their genre and style. To show this we combined all CoSyCo collections of general news into one and calculated all co-occurrence frequencies for the resulting "average" collection. The same was done for IT news. This gave us a chance to compare frequencies in the "average" collection and by each source separately.

In the tables below we show data for several of the semantic classes with the highest scores (which also differed the most).

In each table we included several most frequent nouns and also figures for the whole class in the column Total.

---

[15] Strictly speaking, nouns in the first group may also be used with *gordy* in the second meaning (*gordy*2): e.g. a musician may be called proud not only because of something done out of pride as a character trait, but also because of the emotion felt after the performance. It is also possible (but less typical) to think of an owner of X who treats others with contempt. So it will be more precise to say that nouns in the second group **tend** to be used with *gordy*2 and are separated from all the rest on this basis.

For each word and group "a" columns contain absolute co-occurrence figures for the combination of the adjective *gordy* with this noun or this semantic class of nouns; "b" columns show how often this pair is found as compared to the total number of such pairs in the collection for this source; "c" columns contain relative frequency of the pair occurrence in the "average" collection (for CoSyCo data).

Of the 17 nouns denoting various representational objects (denoting the kinds of names) the following three were most frequent:

	***imya* 'name'**			***zvanie* 'rank'**			***nazvanie* 'name'**			**total**		
	a	**b, %**	**c, %**	**a**	**b, %**	**c, %**	**a**	**b, %**	**c, %**	**a**	**b, %**	**c, %**
RNC 1000	14	3.91		5	1.40		10	2.79		31	8.7	
RNC news	**60**	**5.89**		**42**	**4.12**		**32**	**3.14**		**149**	**14.7**	
GICR news	**39**	**4.63**		**47**	**5.58**		**27**	**3.21**		**126**	**15**	
lenta	1	1.18	0.12	4	4.71	0.47	5	5.88	0.59	15	17.65	1.76
rian	13	4.71	1.53	13	4.71	1.53	11	3.99	1.29	39	14.13	4.58
regnum	18	6.72	2.11	12	4.48	1.41	3	1.12	0.35	37	13.81	4.34
rosbalt	7	3.14	0.82	18	8.07	2.11	8	3.59	0.94	35	15.70	4.11
CoSyCo news												
commersant	14	5.49	0.70	19	7.45	0.95	9	3.53	0.45	47	18.4	2.35
rian	28	4.03	1.40	18	2.59	0.90	21	3.03	1.05	73	10.5	3.64
rg	14	6.06	0.70	11	4.76	0.55	8	3.46	0.40	34	14.7	1.70
utro	3	1.96	0.15	9	5.88	0.45	13	8.50	0.65	26	17	1.30
vz	5	2.75	0.25	14	7.69	0.70	10	5.49	0.50	33	18.1	1.65
nezavisimaya	14	6.36	0.70	8	3.64	0.40	6	2.73	0.30	29	13.5	1.45
lenta	1	1.89	0.05	—	—	—	1	1.89	0.05	2	3.8	0.10
polit	1	1.28	0.05	—	—	—	2	2.56	0.10	4	5.2	0.20
rbc	2	2.13	0.10	6	6.38	0.30	10	10.64	0.50	18	20.2	0.90
ibusiness	3	6.82	0.15	1	2.27	0.05	5	11.36	0.25	9	22.5	0.45
AVG CoSyCo NEWS	**85**		**4.24**	**86**		**4.29**	**85**		**4.24**	**275**		**13.72**
IT news												
cnews	2	9.09	0.48	—	—	—				2	9.09	0.48
compulenta	6	13.95	1.45	1	2.33	0.24	3	6.98	0.72	11	23.26	2.66
computerra	39	14.18	9.42	13	4.73	3.14	21	7.64	5.07	87	26.55	21.01
membrana	3	9.68	0.72	1	3.23	0.24	1	3.23	0.24	7	16.14	1.69
Pc week	2	4.65	0.48	1	2.33	0.24	3	6.98	0.72	6	13.96	1.45
AVG IT NEWS	**52**		**12.56**	**16**		**3.86**	**28**		**6.76**	**113**		**27.29**

It is clear that the weight of this group varies a lot: from 3.8% for Lenta to 26.55% for Computerra. RNC and GICR as non-segregated sources also show rather high scores of 14.7% and 15% (as compared to scores below 5% for most other semantic groups and classes). GICR figures by source vary from 13.81% up to 17.65%. The difference in two Lenta.ru collections stands out. GICR contains a bigger version of this site including both news and analytics, while CoSyCo version contains news wire only. Thus, the difference between figures could be explained by the difference in the collections' style. It is also evident that IT news tend to take the higher end of the range, whereas sources with more varied topics score lower. The reason behind this may be that in the first group the necessity to name new objects (projects, organizations,

etc) is higher. It should be noted that in such cases the word tends to be used in a humourous or ironic way[16].

Of the 24 nouns from another subgroup of representational objects (referring to a way of showing the name/class of the object) the word *nadpis'* 'inscription' stood out from the rest. For this subgroup the figures are much lower[17] than for the *imya* group, but the tendency still remains for IT news to occupy the upper end of the diapason.

Of the 33 words grouped as names of (emotional) states, feelings, personality traits two words show high frequencies, far outstepping all the rest:

	***odinochestvo* 'loneliness'**			***molchanie* 'silence'**			**total**		
	a	**b**	**c**	**a**	**b**	**c**	**a**	**b**	**C**
RNC 1000	32	8.94		5	1.40		51	14.2	
RNC news	**174**	**17.08**		**12**	**1.18**		**195**	**19.1**	
GICR news	**137**	**16.27**		**22**	**2.61**		**168**	**20**	
lenta	12	14.12	1.41	2	2.35	0.23	14	16.47	1.64
ria	52	18.84	6.10	1	0.36	0.12	57	20.65	6.69
regnum	42	15.67	4.93	14	5.22	1.64	59	22.01	6.92
rosbalt	32	14.35	3.76	5	2.24	0.59	38	17.04	4.46
CoSyCo news									
commersant	35	13.73	1.75	2	0.78	0.10	41	16.1	2.05
rian	73	10.52	3.64	3	0.43	0.15	81	11.7	4.04
rg	43	18.61	2.15	3	1.30	0.15	56	24.2	2.79
utro	38	24.84	1.90	2	1.31	0.10	40	26.1	2.00
vz	28	15.38	1.40	1	0.55	0.05	30	16.5	1.50
nezavisimaya	34	15.45	1.70	4	1.82	0.20	44	20.5	2.20
lenta	—	—	—	—	—	—	—	—	0.00
polit	8	10.26	0.40	1	1.28	0.05	11	14.1	0.55
rbc	24	25.53	1.20	—	—	—	24	25.5	1.20
ibusiness	3	6.82	0.15	—	—	—	3	7.5	0.15
AVG CoSyCo NEWS	**286**		**14.27**	**16**		**0.80**	**330**		**16.47**
IT news									
cnews	2	9.09	0.48	1	4.55	0.24	3	13.6	0.72
compulenta	14	32.56	3.38	—	—	—	14	32.6	3.38

[16] Irony and humour are considered examples of the so-called *non bona fide* modus of discourse (Shilikhina 2014). They appear when there is intended incoherence in the utterance (i.e. a disruption in semantic cohesion within the utterance or an incongruity between the utterance and the situation described), which signals the presence of implicit meanings. Irony presupposes implicit negative deontic assessment.

IT news attracted our attention in this respect: when we checked the proportion of humourous and ironic contexts in IT news, we managed to find about one or two "serious" (bona fide) uses of *gordy* per 100 sentences. This figure for general news texts is lower (varies in the range of 10–30%), but is still presumably much higher than in fiction (this requires further research).

[17] For this reason we will not include these data here.

	odinochestvo 'loneliness'			*molchanie* 'silence'			total		
	a	b	c	a	b	c	a	b	C
computerra	29	10.55	7.00	6	2.18	1.45	35	12.73	8.45
membrana	7	22.58	1.69	—	—	—	7	22.6	1.69
Pc week	13	30.23	3.14	1	2.33	0.24	14	32.6	3.38
AVG IT NEWS	**65**		**15.70**	**8**		**1.93**	**73**		**17.63**

Odinochestvo (which accounts for 60–90% of the class weight) also shows remarkable variation from 7.5% (ibusiness) up to 32.56% (computerra), with IT news once again taking the higher end of the range. *Molchanie* shows the same tendency, though with lower figures. In polythematic sources *gordy* is more often used with other nouns from this group. Both these words show such high frequencies because they are used in clichéd expressions as desemantised phrasemes marking humour or irony in the utterance. Here the difference between the two versions of Lenta.ru becomes even more obvious: if these markers of humour and irony are not found in "normal" news at all, it is definitely not so for analytics (the figures are on a par with other polythematic sources).

The group of nouns naming a person according to nationality/ethnicity is remarkable in the sense that of the 86 nouns belonging to the group it is hard to name any which would be used more often than the rest (let alone do it consistently through several sources). The total class figures (which will be not given here for lack of space) show that the class accounts for more than 5% of *gordy* usage for many sources, and predictably the scores are higher for polythematic sources.

Of the 24 nouns referring to a big group of people (community) the following three were more frequent:

	narod 'a people'			*strana* 'a country'			*gosudarstvo* 'a state'			total		
	a	b, %	c, %	a	b, %	c	a	b, %	c, %	a	b, %	c, %
RNC 1000	5	1.40		4	1.12		—	—		9	2.5	
RNC news	**15**	**1.47**		**16**	**1.57**		**5**	**0.49**		**57**	**5.6**	
GICR news	**61**	**7.24**		**21**	**2.49**		**15**	**1.78**		**127**	**15.1**	
lenta	3	3.53	0.35	1	1.18	0.12	—	—	—	9	10.59	1.06
ria	17	6.16	2.00	3	1.09	0.35	2	0.72	0.23	26	9.42	3.05
regnum	26	9.70	3.05	11	4.10	1.29	7	2.61	0.82	59	22.01	6.92
rosbalt	15	6.73	1.76	6	2.69	0.70	6	2.69	0.70	33	14.80	3.87
CoSyCo news												
commersant	5	1.96	0.25	9	3.53	0.45	1	0.39	0.05	23	9.1	1.15
rian	25	3.60	1.25	11	1.59	0.55	2	0.29	0.10	48	6.9	2.40
rg	6	2.60	0.30	2	0.87	0.10	1	0.43	0.05	13	5.6	0.65
utro	10	6.54	0.50	3	1.96	0.15	1	0.65	0.05	18	11.8	0.90
vz	2	1.10	0.10	2	1.10	0.10	1	0.55	0.05	9	4.9	0.45
nezavisimaya	5	2.27	0.25	1	0.45	0.05	—	—	—	13	6	0.65

	narod 'a people'			*strana* 'a country'			*gosudarstvo* 'a state'			total		
	a	**b, %**	**c, %**	**a**	**b, %**	**c**	**a**	**b, %**	**c, %**	**a**	**b, %**	**c, %**
lenta	1	1.89	0.05	—	—	—	—	—	—	2	3.8	0.10
polit	5	6.41	0.25	3	3.85	0.15	1	1.28	0.05	12	15.4	0.60
rbc	1	1.06	0.05	4	4.26	0.20	—	—	—	5	5.3	0.25
ibusiness	—	—	—	1	2.27	0.05	—	—	—	2	5	0.10
AVG CoSyCo NEWS	**60**		**2.99**	**36**		**1.80**	**7**		**0.35**	**145**		**7.24**
IT news												
cnews	—	—	—	4	18.18	0.97	—	—	—	4	18.18	0.97
compulenta	—	—	—	—	—	—	—	—	—	—	—	—
computerra	2	0.73	0.48	2	0.73	0.48	—	—	—	8	2.92	1.92
membrana	—	—	—	—	—	—	—	—	—	—	—	—
Pc week	—	—	—	—	—	—	—	—	—	—	—	—
AVG IT NEWS	**2**		**0.48**	**6**		**1.45**	**—**		**—**	**12**		**1.93**

Results predictably show that scores are higher for polythematic sources, especially for those more focused on politics.

5. Conclusion

As we can see, figures for "generalized" corpora change more or less in the same way, accurate to the selection of chosen sources. However, splitting the corpus into subcorpora leads to significant changes in the word usage distribution. Moreover, stylistically different parts of the same corpus show dramatic differences. IT news sources show a similar tendency.

Results demonstrate that at the first glance in general the usage of *gordy* in news and newspaper texts can be as varied as in fiction: most of the classes identified in fiction are present in many sources. However, it is clear that *gordy* tends to appear more frequently with nouns in several particular zones from the whole list of possible combinations. Such classes as nouns naming a person according to nationality/ethnicity, big groups of people or representational objects and desemantised phrasemes are prominent for polythematic news sources. For IT news names of representational objects and desemantised phrasemes are the most frequently used, scoring higher than the same classes in general news.

The choice of such zones predictably depends on the topics covered by the source. Another crucial factor is the style of the source. The more markers of humour and irony (Shilikhina 2014) in the source, the more probable it is that *gordy* is used not seriously and expresses negative attitude towards the object or event characterized as such.

In sum, it may seem obvious that if we combine together texts of different styles and genres we need to be able to study them separately. Otherwise if we study them as if they were a homogeneous text, we get results which conceal the existing genre and style features. Existing corpora are gradually becoming more flexible in this respect, as they start to allow separating the data from different sources. It remains

an open question what particular styles and genres should be included in a corpus which is intended as suitable for various kinds of research. What degree of granularity it would be reasonable to ensure in such a corpus is also a matter of further studies.

References

1. *Belikov V., Selegey V., Sharoff S.* (2012), Preliminary considerations towards developing the General Internet Corpus of Russian [Prolegomeny k proektu General'nogo internet-korpusa russkogo yazyka], Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog" 2012" [Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Po materialam ezhegodnoi Mezhdunarodnoii Konferentsii "Dialog 2012"], Bekasovo, vol. 1, pp. 37–49.
2. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.* (2013), Corpus as language: from scalability to register variation [Korpus kak yazyk: ot masshtabiruemosti k differentsialnoi polnote] Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialog" (2013) [Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Po materialam ezhegodnoi Mezhdunarodnoii Konferentsii "Dialog" (2013)], Bekasovo, vol. 1, pp. 83–96.
3. *Kilgarriff A., Rychly P., Smrz P., Tugwell D.* (2004), The Sketch Engine, Proceedings of the XI Euralex International Congress, Lorient, France, pp. 105–116.
4. *Klyshinsky E., Kochetkova N., Litvinov M., Maximov V.* (2011), Method of POS-disambiguation using information about words co-occurrence (for Russian), Proceedings of the annual meeting of the Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL), Hamburg, pp. 191–195.
5. *Lashevskaja, O., Plungian V.* (2003), Morphological annotation in Russian National Corpus: a theoretical feedback, Proceedings of the 5th International Conference on Formal Description of Slavic Languages (FDSL-5), Leipzig, pp. 26–28.
6. *Shilikhina K. M.* (2014), Semantics and Pragmatics of Verbal Irony [Semantika i pragmatika verbal'noi ironii], NAUKA-UNIPRESS, Voronezh.
7. *Vlasova A. A., Korolyov D. V., Klyshinsky E. S.* (2016), Development of the software tool for searching of clauses in untagged texts [Razrabotka instrumental'nogo sredstva dlia poiska sintaksicheskikh konstruktsij v nerazmechennoj kollektsii tekstov], Proceedings of New Information Technologies in Automated Systems, Ekaterinburg, pp. 81–84.

WELCOME TO THE CLUB: DESIGNING THE INVENTORY OF SEMANTIC ROLES FOR ADJECTIVES[1]

Lyashevskaya O. N. (olesar@yandex.ru)
National Research University Higher School of Economics,
Vinogradov Institute of the Russian Language RAS,
Moscow, Russia

Kashkin E. V. (egorkashkin@rambler.ru)
Vinogradov Institute of the Russian Language RAS,
Moscow, Russia

The argument constructions of adjectives has largely been out of the scope of research on semantic roles both in theoretical and IT fields. Before adding the roles of adjectival arguments to the network of semantic roles it is important to determine whether the adjectival roles form a separate list or whether they can be seen as an extension of roles assigned to the patterns of verbs and nominalizations. We discuss the general principles of how the inventory of adjectival roles should be organized in comparison with the existing inventories of verbal roles. In order to verify our statements, we carry out an experimental survey aimed at measuring the similarity between adjectival and verbal roles. The results have shown that both semantic interpretation of roles and their typical morpho-syntactic expression are significant for the evaluation and should be taken into account in working out the inventory. Besides, the specificity of adjectives lies in their prototypical stative semantics, which favors some differences in assigning a semantic role as compared to verbs. The results of the survey also provide some evidence for verification and development the inventory of verbal semantic roles.

Keywords: semantic roles, semantic similarity, predicate-argument constructions, adjectives, verbs, Russian language, experimental linguistics, inter-rater agreement

[1] The work was partly supported by the Russian Basic Research Foundation, grant No. 15-07-09306.

О МЕСТЕ СЕМАНТИЧЕСКИХ РОЛЕЙ ИМЕН ПРИЛАГАТЕЛЬНЫХ В ГРАФЕ РОЛЕЙ

Ляшевская О. Н. (olesar@yandex.ru)

Национальный исследовательский университет Высшая школа экономики; Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Кашкин Е. В. (egorkashkin@rambler.ru)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Предикатно-аргументные конструкции имен прилагательных, в отличие от глагольных конструкций, чаще всего остаются вне зоны внимания как теоретиков, так и специалистов IT-отрасли. Ставя вопрос о включении семантических ролей прилагательных в общую сеть семантических ролей, прежде всего, важно определиться, образуют ли они отдельную систему или их можно рассматривать как расширение инвентаря ролей глаголов и номинализаций. Для проверки наших предположений о принципиальном устройстве системы адъективных ролей мы провели опрос экспертов, в котором просили оценить сходство между ролями прилагательных и глаголов. Результаты показали, что и семантическая интерпретация ролей, и их морфо-синтаксическое оформление оказывают влияние на оценку, а следовательно, должны быть приняты во внимание при разработке инвентаря. Кроме того, прототипически имена прилагательные имеют стативную семантику, и это находит отражение в том, насколько близкими воспринимаются роли участников при прилагательном и глаголе. Результаты опроса дают также новые данные для проверки и уточнения инвентаря семантических ролей самих глаголов.

Ключевые слова: семантические роли, семантическая близость, предикатно-аргументные конструкции, модель управления, имя прилагательное, глагол, русский язык, экспериментальное исследование, согласие асессоров

1. A new species or an extension to the known network?

The classification of semantic roles is an important issue in both theoretical and computational tasks. The theoretical notion of a semantic role contributes to the study of the semantic-syntax interface, for example, in explaining which semantic differences between the arguments interfere with the differences in their morpho-syntactic marking. In computational linguistics, this concept lies at the foundation of semantic role labeling (Màrquez et al. 2008, Palmer et al. 2013, Kuznetsov 2015) and other fields which involve natural language understanding. The problem is, however, that manually created lists of semantic roles (see e.g. Fillmore 1968; Berkeley FrameNet,

Dowty 1991, Apresjan 1995: 125–126; Apresjan et al. 2010: 370–377, Paducheva 2004: 587–588) are sometimes fundamentally different. They vary greatly in number and in the ways particular roles can be interpreted, cf. for example the narrow inventory suggested in [Fillmore 1968] and the potentially unlimited inventory of Berkeley FrameNet including such roles as Agriculturist, Colonists, Electricity etc.

The issue of how the inventory of semantic roles should be designed has been posed primarily for verbal arguments (including nominalized patterns). As regards the arguments of concrete nouns (e.g. *dyra* ***v polu*** 'hole in the floor', *kofe* ***s molokom*** 'coffee with milk') and adjectives (e.g. *dal'ekij* ***ot Moskvy*** 'distant from Moscow', *izvesten* ***svoimi publikacijami*** 'famous for their publications', *nepravil'nyj* ***nomer*** 'wrong number'), their classification is hardly elaborated (cf. a few noteworthy remarks in Bulygina, Shmelev 1997: 58–73; Vol'f 1978; Apresjan 2004). To say more, the very idea of adjectives evoking the semantic predicate-argument relations is not generally acknowledged in the computational linguistics community. For example, the Russian semantic analyzer ABBYY Compreno [Anisimovich et al. 2012] considers a noun as a predicate and an adjective as an argument (with the semantic role Property) in attributive constructions, and not vice versa, mostly overlooking the predicative uses of adjectives. PropBank/Ontonotes 5 [Palmer et al. 2005] covers only a limited number of adjectival argument patterns under the following rationale: "Crosslinguistically, it is common for there to be overlap between what is expressed as a verb and what is expressed as an adjective. <...> Because PropBank is in part a resource for machine translation and several parallel PropBanks exist in different languages, it is important to annotate predicate adjectives in English" [Bonial et al. 2015: 59].

The problem arises that the building the inventory of semantic roles for Russian adjectives is a long way behind the current research on Russian adjectives, cf. [Arkhangelskiy et al. 2010; Kustova 2007, 2009; Rakhilina et al. 2010, among others]. There are no full inventories of semantic roles for adjectives which could be accepted as a gold standard or at least as a starting point. Neither can we rely on any SRL system developed for Russian (since they are still at an early stage, cf. Kuznetsov 2015, Shelmanov, Smirnov 2014) or any other language and obtain objective evaluation metrics for different inventories of semantic roles. Since there is no established tradition of labeling the semantic roles of adjectival arguments, it is important to form the opinion of the community taking into account possible divergences. Rather than building a theory from scratch, we propose a bottom-up experimental approach based on experts' judgement.

In this paper, we probe the hypothesis that the roles in adjectival patterns are (at least to some extent) congruent to the roles of verbal arguments. We suggest that empirical evidence gathered in an experiment while collecting experts' judgements will reveal certain implicit knowledge and assumptions on how the pattens are structured and what priorities the researchers have regarding links between them. In Section 2, we present the design of the experiment and the principles for selecting the verbal roles stimuli. Section 3 outlines the results of the experiment: we discuss here what factors have proved to be relevant for creating the inventory of roles for adjectives. In Section 4, we analyze some evidence provided by our survey which could be helpful for improving the inventory of verbal roles. Section 5 concludes.

2. Questionnaire and data

2.1. Design of the experiment

20 adult native Russian speakers (mean age 34, sd = 16) participated in the study. All respondents were either students of linguistics or professional linguists (lecturers, researchers, developers of computational linguistic systems), which presupposed that they were acquainted with at least one of the existing inventories of verb roles. The questionnaire was anonymous (only the sociolinguistic data on occupation and age were collected) even though the participants could optionally provide their name and email address if they were interested in feedback regarding the results of the survey. The survey was administered as an online questionnaire with no time limits. The expected time for its completion was 20–30 minutes.

The experiment was designed as a score-assignment test. The participants were asked to rate the similarity between the target pair ADJECTIVE—ITS ARGUMENT and the control pair VERB—ITS ARGUMENT according to a scale of 1 to 7, see Fig. 1. The stimuli included 16 target sentences which illustrated the use of seven adjectives (*gotovyj* 'ready', *svobodnyj* 'free (from)', *sil'nyj* 'strong, impressive, *blizkij* 'close', *ščedryj* 'generous', *izvestnyj* 'famous for, known by', *vinovatyj* 'guilty') in different meanings and in different morphosyntactic patterns (e.g. with different dependent prepositional phrases, see the examples in Sections 3 and 4). The difference between attributive and predicative uses was not specially investigated in this study, the examples included both types of syntactic patterns (10 predicative constructions among the 16 target entries and 6 attributive constructions).

Some adjectives in our smaple have cognate verbs (e.g. *gotovyj—gotovit'*, *blizkij—priblizit'*, *vinovatyj—obvinit')*. In some cases we might think of simply transferring verbal arguments and their roles to adjectival constructions. However, this decision is not applicable in the general case due to the possible asymmetry between verbal and adjectival valency patterns, cf. *Ja* ***gotov*** *pomoč' tebe* 'I am ready to help you' vs. **On* ***gotovit*** *men'a pomoč' tebe*, expected meaning 'He is making me ready to help you'; *Ja* ***vinovat*** *pered Vami* 'I am guilty towards you' vs. **On* ***obvinil*** *men'a pered Vami*, lit. 'He accused me towards you'.

Each target sentence was followed by 3–4 control pairs VERB—ITS ARGUMENT also shown in a sentence. The pairs had been selected in such a way that they would range from very similar to hardly similar to the target adjectival pair (according to preliminary judgements of the authors and taking into account both their frame semantics and morphosyntax). It was possible for the participants (but not obligatory) to suggest their own version of the pair VERB—ITS ARGUMENT most close to the target stimulus (a free answer field in the questionnaire). However, we have not received free answers indicating that any variants which could possibly gain a high score were missing.

Each participant went through all 58 questions. The order of adjectives and questions was randomized in four sets of stimuli. A sample questionnaire is available at https://goo.gl/xy8ST0.

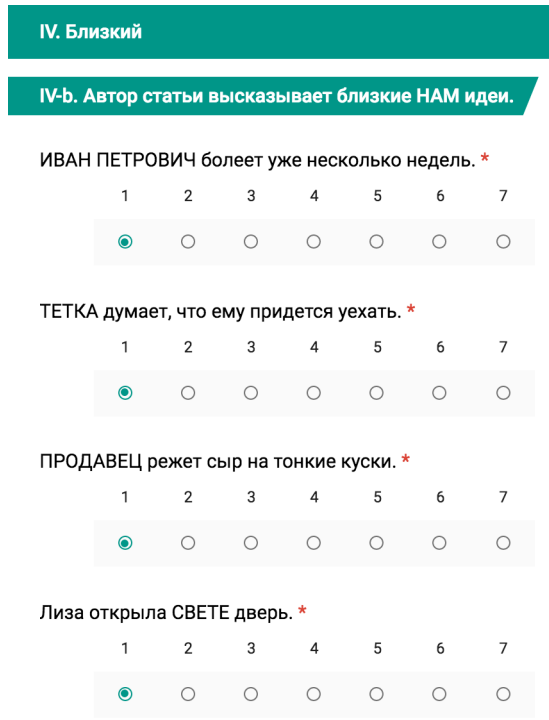


Fig. 1. A block of questions. The adjective *blizkij* 'close' in the target context *Avtor stat'ji vyskazyvaet blizkie NAM idei* 'The author of the article puts forward ideas close TO US' and four contexts to be evaluated: *IVAN PETROVIČ boleet uzhe neskol'ko nedel'* 'IVAN PETROVICH has been sick for a few weeks', *TETKA dumaet, čto emu pridets'a ujehat'* 'His AUNT thinks that he will have to leave', *PRODAVETS režet syr na tonkie kuski* 'The SELLER slices the cheese into thin slices', and *Liza otkryla SVETE dver'* 'Liza opened the door FOR SVETA'.

2.2. The inventory of verbal roles

Our research relies on Russian FrameBank (http://www.framebank.ru). This is an open access database which includes a dictionary of Russian lexical constructions and a corpus of their uses tagged with a FrameNet-like annotation scheme (see [Lyashevskaya 2010; Lyashevskaya, Kashkin 2015a, b] for details). At present the dictionary provides data for ca. 4,000 target verbs, adjectives, and nouns, and the corpus part includes ca. 50,000 annotated examples. Constructions of each verb in the dictionary differ, first, in the morpho-syntactic pattern, and, second, in the meaning of a verb.

FrameBank includes the elaborated inventory of verbal roles, which we can rely on in our study. Bearing in mind the major differences between different role inventories and some vagueness of the task to create a role inventory good for all purposes, we aim to develop the inventory for adjectives within the existing annotation scheme of FrameBank. The inventory of semantic roles used in FrameBank contains 91 roles and is based on the following principles (for a more detailed discussion see [Kashkin, Lyashevskaya 2013], [Lyashevskaya, Kashkin 2015b]):

- The roles correlate with the semantic classification of the lexicon. Traditionally "broad" roles such as Agent or Patient should get different labels in different semantic classes, cf. Agent in destruction vs. speech vs. motion
- The roles of semantically close lexemes should systematically coincide or systematically differ.
- The full inventory of roles should cover all the lexical domains.
- The inventory is organized hierarchically in order to provide flexible search options (see the role network at http://marker.framebank.ru/GraphSemRoles.pdf).
- The scope of a semantic role follows the principle of a prototype and its periphery. For instance, the prototype of Patient is a participant changing under the physical influence of an Agent; peripheral examples (Patient of a non-physical process, Patient which is not changing, Patient created as a result of a physical action) get specific labels (Theme, Result, etc.) and are considered as specific types of Patient.

However, the database of FrameBank includes primarily verbs, whereas an adequate sample of adjectives is still to be added there. The inventory of semantic roles for adjectives has not been fully developed either. In order to get some verifiable evidence on how this inventory should be organized, we have carried out a survey using some data on verbal roles implemented into the dictionary of FrameBank.

3. Analysis

3.1. From roles of verbal arguments to roles of adjectival arguments

Figure 2 summarizes the results of the survey. Each bar represents the mean score of each question, the vertical line above and below the bar being the standard deviation of individual scores. For each adjectival role, the results are ordered from the best matching verbal role to the poorest matching one. The multi-rater agreement was, predictably, not very high (exact Conger's kappa = 0.0579, light kappa = 0.0604[1][2]) since the scores were subjective and based on different theoretical assumptions on how semantic roles are classified: some of the respondents draw subtle semantic distinctions typical of fine-grained role inventories, whereas others may combine rather heterogenous entities within one class.

[2] The scores are obtained using the function KappaM in DeskTools package of R.

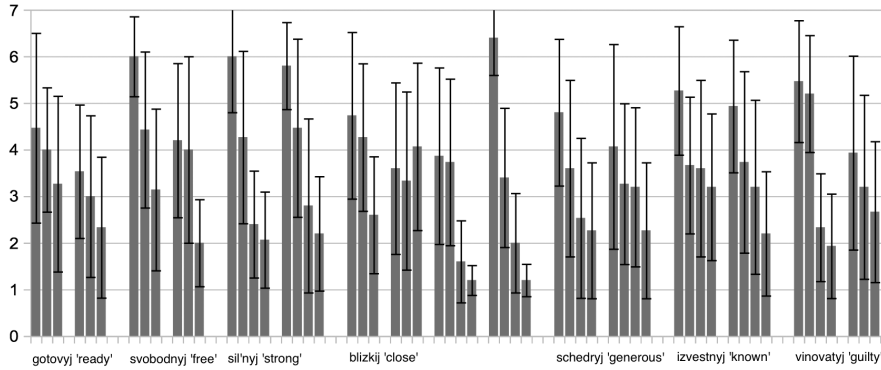


Fig. 2. Rater scores for the questionnaire

On the whole, the results suggest that the roles of adjectival arguments (at least provided in our data sample) can be adequately tagged using the inventory of roles describing verbal arguments. 6 of the 16 experimental blocks contain an example which has received a high average mark from 5 to 7 showing the similarity between the roles of the target verbal argument and of the target adjective. These examples are listed below (the role of the verbal argument in FrameBank and the average mark are given in brackets after an example):

(1) *Natal'ja Jur'jevna byla očen' blizka* ***s otcom*** 'Natalya Jurievna was very close to her father'—*Kol'a družit* ***s Natašej*** 'Kolya is friends with Natasha' (Counter-Agent of social relation; 6.4)

(2) *Avar'ijnye vyhody i prohody dolžny byt' svobodny* ***ot ručnoj kladi*** 'Emergency exits and passages must be free of hand luggage'—*My očistili čerdak* ***ot hlama*** 'We cleared the attic of junk' (Patient; 6).

(3) *Pet'a sil'en* ***v matematike*** 'Petya is good (lit.: strong) at mathematics'—*On vseh obošel* ***v učebe*** 'He left everyone behind in his studies' (Sphere; 6).

(4) *Zhdanov sil'en* ***ritorikoj*** 'Zhdanov is impressive (lit.: strong) in his rhetoric'—*Sredi sverstnikov on vydel'als'a* ***svoej l'uboznat'el'nost'ju*** 'He was notable among his peers for his curiosity' (Quality; 5.8).

(5) *Ja vinovat* ***pered vami*** 'I am to blame for doing something wrong to you (lit.: guilty towards you)'—*On ne stanet unižat's'a* ***pered načal'nikom*** 'He will not humiliate himself in front of his boss' (Counter-Agent of social relation; 5.5), *On izvinils'a* ***pered passažirami*** 'He apologized to the passengers' (addressee; 5.2).

(6) *Etot žurnalist izvesten* ***svoimi razoblačitel'nymi publikacijami*** 'This journalist is famous for his unmasking publications'—*Pavel porazil vseh* ***dlinnymi volosami*** 'Pavel amazed everyone with his long hair' (Property of the Reason for emotional state; 5.2)

Among the other 10 blocks, 7 blocks include at least one argument with the average score from 4 to 5, for instance:

(7) *Samymi* *ščedrymi* ***na novogodnie podarki*** *v etom godu stanut rukovoditeli rossijskih kompanij i gosslužaščie* 'This year the most generous in giving New Year's presents will be the managers of Russian companies and civil servants'—*Deduška* *dal* *rebenku* ***konfetu*** 'Grandfather gave a sweet to a child' (Patient; 4.8).

(8) *Ja* *gotov* ***pomoč tebe*** 'I am ready to help you'—*My* *hotim* ***popast' na vystavku*** 'We want to get to the exhibition' (Content of thought; 4.5), *I vot nastupil den', kogda ranenyj* *smog* ***vstat'*** 'Finally the day came when the wounded man was able to stand up' (Content of action; 4).

A factor that might have influenced some answers of the respondents is whether an adjective is used in an attributive construction or in a predicative one: the latter may be expected to be more "verbal". As can be seen from the examples above, there are many predicative uses with high evaluation. However, some attributive constructions have also gained high scores, e.g. (7). Furthermore, we have received plenty of low scores for the predicative constructions, i. e. this type of syntactic construction is not necessarily evaluated as semantically similar to a random verbal construction, and the survey shows substantial differences in comparing one and the same predicative use of an adjective with verbal constructions varying in their role pattern. This is what we have actually expected to test, and in this sense the opposition between attributive and predicative uses does not interfere with our conclusions. A further interesting point could lie in comparing evaluations for attributive and predicative uses of one and the same adjective, but this task has so far remained beyond the scope of our research.

Our survey has therefore provided a representative subset of the role inventory for adjectives. These roles come from the verbal role inventory. Their list can be found in Table 1. The roles are provided with examples of adjectives taken from the experimental data. The morphosyntactic constructions are labelled according to the general annotation scheme of Russian FrameBank (where Sx means 'substantive in the case x').

In total, Table 1 includes 14 roles with the overall rating 4 or higher. Surely, this list is not exhaustive, as we have not aimed at creating its final version, and the experimental data does not cover all possible adjectival constructions. Rather, we put forward the hypothesis that verbal roles can be transferred to adjectival constructions and confirmed it by our experiment. The inventory from Table 1 can be enlarged following the principles which result from our survey and which will be discussed in the next sections.

Table 1. Inventory of semantic roles for adjectives: a preliminary list for the study

Semantic role	Adjective and morpho-syntactic pattern
Counter-Agent of social relation	*blizkij s* + S_{ins} 'close to smb (e.g., a friend)', *vinovatyj pered* + S_{ins} 'guilty towards smb'
Patient	*svobodnyj ot* + S_{gen} 'free from sth', *schedryj na* + S_{acc} 'generous in sth (e.g., gifts)'
Content of thought	*gotovyj* V_{inf} 'ready to do sth.'
Content of action	*gotovyj* V_{inf} 'ready to do sth.' (a competing role in the results of the survey)
Content of utterance	*schedryj na* + S_{acc} 'lavish with sth (speech etc.)'
Addressee	*vinovatyj pered* + S_{ins} 'guilty towards smb' (a competing role in the results of the survey)
Beneficiary	*svobodnyj dl'a* + S_{gen} 'free for smb/sth'
Location	*blizkij k* + S_{dat} / *ot* + S_{dat} 'close to sth.'
Point of destination	*blizkij k* + S_{dat} 'close to sth.' (a competing role in the results of the survey)
Sphere	*sil'nyj v* + S_{loc} 'strong in sth.'
Social environment	*izvestnyj v* + S_{loc} 'famous among smb'
Goal	*svobodnyj dl'a* + S_{gen} 'free for smb/sth' (a competing role in the results of the survey)
Feature	*sil'nyj* S_{ins} 'strong with sth.'
Property of Reason for mental state	*izvestnyj* S_{ins} 'famous with sth.'

3.2. Roles of adjectival arguments: semantics vs. morpho-syntax

Let us now go on to the possible principles for assigning roles to adjectival arguments. As can be seen from Section 3.1, adjectival arguments with high average scores often take the same morpho-syntactic marking as the corresponding verbal arguments. However, this is not always the case, cf. example (7) where a Given thing is involved into different constructions with the verb *dat'* 'to give' and the adjective *ščedryj* 'generous'. On the whole, the factors of semantics and morpho-syntax interact in assigning semantic roles to adjectival arguments in our data[3]. There are two important trends which follow from our survey.

First, the choice of a semantic role is deeply influenced by the semantic classes of the target verb / adjective and of their target arguments. Verbal constructions with

[3] The coarse binary classification "same VS different" of similarity in meaning and morpho-syntax applied to the data shows a significant difference in the scores obtained in four groups (Chisq. p-value = 0.0009, df = 1).

the same morpho-syntactic marking get the higher score, the closer they are semantically to the adjectival construction from a given experimental block. This supports the idea that the classification of semantic roles should correlate with the semantic classification of verbs suggested in [Kashkin, Lyashevskaya 2013; Lyashevskaya, Kashkin 2015b] and implemented in the dictionary of FrameBank. Thus, the closest verbal construction for the adjectival example *Pet'a sil'en* ***v matematike*** 'Petya is good (lit.: strong) at mathematics' is represented in the sentence *On vseh obošel* ***v učebe*** 'He left everyone behind in his studies' (Sphere; 6), while the same morpho-syntactic construction from *Ja živu* ***v Moskve*** 'I live in Moscow' (Place; 2.4) receives a significantly less average score (2.4 vs. 6). Similarly, the example *Etot žurnalist izvesten* ***svoimi razoblačitel'nymi publikacijami*** 'This journalist is famous for his unmasking publications' has been primarily related to *Pavel porazil vseh* ***dlinnymi volosami*** 'Pavel amazed everyone with his long hair' (Property of the Reason for emotional state; 5.2), whereas the instrumental construction from *Ivan razbil okno* ***palkoj*** 'Ivan broke the window with a stick' gets the much less average score of 3.2 due to the semantic difference between the verbs of mental state and of physical impact.

Second, if an experimental block includes several verbal constructions which can be treated as adequate semantic correspondences to the target entry, the respondents tend to choose the closer morpho-syntactic pattern. For instance, the argument of the adjective *svobodnyj* 'free' in *Avar'ijnye vyhody i prohody dolžny byt' svobodny* ***ot ručnoj kladi*** 'Emergency exits and passages must be free of hand luggage' receives the same role as the prepositional phrase in *My očistili čerdak* ***ot hlama*** 'We cleared the attic of junk' (Patient; 6), while the direct object in *My ubrali* ***al'bomy*** *s polok* 'We removed the albums from the shelves' gets a nearly half the average score (3.1) due to its different syntactic status. The adjective *izvestnyj* 'known, famous' in the sentence *Policija zaderžala narkotorgovca, izvestnogo* ***v opredelennykh krugakh*** *pod kličkoj "Korotyška"* 'The police arrested a drug pushed known as "Shorty" in criminal circles' has an argument which is more probably related to the argument expressing Social circle in *Soobščenie posejalo paniku* ***v r'adah vraga*** 'The message spread panic among the ranks of the enemy' (4.9) than to the Subject of mental state in ***Ivan*** *znaet, čem končilos' delo* 'Ivan knows how the things have finished' (3.2). While the semantics of the highlighted arguments in both verbal constructions is adequate for the target adjective, the most preferable is the example with the same syntactic rank of the argument.

An interesting example of how morpho-syntax and lexical semantics interact in assigning a semantic role is provided by the example *Avtor stat'ji vyskazyvaet blizkie* ***nam*** *idei* 'The author of the article puts forward ideas close to us'. There are two verbal constructions with nearly the same similarity rank here: ***Tetka*** *dumaet, čto emu pridets'a ujehat'* 'His aunt thinks that he will have to leave' (Subject of mental state, 3.7) and *Liza otkryla* ***Svete*** *dver'* 'Liza opened the door for Sveta' (Benefactive, 3.9). The experiencer-like role of Subject of mental state might seem a more precise semantic label for the argument of *blizkij* 'close' in the sentence above, however its syntactic function is different, which might probably have reduced its average rank in our survey. On the contrary, the benefactive argument in 'open the door for X' takes the same syntactic marking and has been evaluated as a more exact correspondence to the argument of *blizkij* 'close', despite its semantic distance from the prototype of a beneficiary.

3.3. Semantics: stative vs. dynamic

One more semantic factor important in working out the inventory of roles for adjectives is static vs. dynamic character of a situation. While verbs show a great variety in their aspectual properties, adjectives prototypically refer to states. According to our study, sometimes this may provoke the difference between role patterns of verbs and adjectives. For example, the adjective *blizkij* 'close' in its literal spatial meaning can bear arguments marked either as Point of destination (9) or as Initial point (10)—the latter class of examples is probably not a prototype for this adjective, but it does occur in the Russian National Corpus and has therefore been included into our survey.

(9) *Ekspedicija obsledovala blizkie* ***k Saransku*** *sela* 'The expedition explored the villages close to Saransk'

(10) *V tu že subbotu, rannim večerom, uspel Aleksandrov sbegat' s kon'kami na neboľšoj, no ujutnyj i blizkij* ***ot doma*** *katok Patriarših prudov* 'On the same Saturday, early in the evening, Aleksandrov had time to run with his skates to the small but cosy skating-rink of Patriarshie ponds, which was close to (lit.: from) his home'

The respondents had to evaluate these examples against verbal constructions with Point of destination (*podojti k domu* 'approach the house'), Initial point (*otojti ot dveri* 'move away from the door'), and Location (*hodit' u reki* 'go along the river'). The results are summarized in Table 2:

Table 2. Role assessment for the arguments of *blizkij* + *k* 'close to' and *blizkij* + *ot* 'lit. close from'

	Point of destination	**Initial point**	**Location**
blizkij k + S_{dat} (Point of destination—like)	4.3	2.6	**4.7**
blizkij ot + S_{dat} (Initial point—like)	3.6	3.3	**4.1**

Table 2 shows that the locative arguments of *blizkij* 'close' are more likely to get the role of location typical of statives like 'to be', 'to live', etc. The intrinsic stative nature of adjectives is therefore the most important factor here. However, the syntactic marking of the argument has also influenced the preferences of our respondents, cf. the values for Point of destination and Initial point. Note also that the results for *blizkij* does not match the argument structure of its verbal cognate *priblizit'* 'to bring nearer', as the latter cannot take a stative argument marked as Location and typical of existential or posture predicates.

4. On benefits for the structure of verbal roles

The results of our experiment highlight some points in how the inventory of verbal roles is organized. As has already been mentioned, the roles used in the Russian FrameBank are structured as a network. The graph was created manually based on semantic similarity between the roles (see the definitions in [Lyashevskaya, Kashkin 2015b: 505–525] and the references therein). The judgments obtained from our survey help to verify the decisions we have previously taken. We rely on the following principle here. If two examples with verbs both get a high average score of their semantic similarity to a given adjectival construction, the roles of target verbal arguments in these examples are also evaluated as semantically similar. If two examples with verbs get significantly different scores, the target roles are also assessed as considerably different. If both verbal examples get low scores, it means nothing for comparing the two verbal roles, since they can diverge from the target adjectival role in different ways.

A case study can be provided by the following experimental block:

(11) *Natal'ja Jur'jevna byla očen' blizka* ***s otcom*** 'Natalya Jurievna was very close to her father'

(12a) ***On*** *vospityvaet trjoh synovej* 'He brings up three sons' (Subject of social relation; 2)

(12b) *Kol'a družit* ***s Natašej*** 'Kolya is friends with Natasha' (Counter-Agent of social relation; 6.4)

(12c) *Mit'a podrals'a* ***s Lešej*** 'Mitya fought with Lyosha' (Counter-Agent; 3.4)

(12d) ***Krest'janin*** *rubit drova* 'The peasant is chopping firewood' (Agent; 1.2)

The graph suggested in [Kashkin, Lyashevskaya 2013] represents Counter-Agent and Subject of social relation as subtypes of Agent, whereas Counter-Agent of social relation is considered a subtype of both Counter-Agent and Subject of social relation. While the latter decision does not prove to be inadequate in our experimental data (Counter-Agent receives a not quite low score), Agent and Subject of social relation have got low scores, which present a challenge for further refinements of a role hierarchy.

5. Conclusions

The survey allowed us to formulate some principles which could govern assigning semantic roles to adjectival arguments. We have shown that adjectives and verbs can share the same role inventory, since quite a few verbal roles were evaluated as good candidates for adjectival constructions. The inventory of adjectival roles can be at least a subset of the inventory intended for verbs (however, we cannot infer from our survey

whether two role inventories are all the same, because the experimental data was still limited). based on the inventory of verbal roles. On the whole, the choice should be based on the semantic similarity between adjectival and verbal arguments. However, if there are several variants possible on semantic grounds, the following principles come into force:

- The priority should be given to a role which is expressed by a verbal argument of the same syntactic rank as the target adjectival argument.
- If there are several possible candidates belonging to either stative or dynamic verbs, the priority should be received by the roles of stative verbs, due to the prototypically stative nature of adjectives.

The experiment allowed us to produce the first draft of the role inventory for adjectival constructions. Including 14 items at present, it will obviously be enlarged at the next research steps following the principles discussed in this paper.

A further step of our project will consist of implementing this strategy into the dictionary of adjectival constructions in FrameBank, together with elaborating it for a bigger data set with more adjectives and more subtle semantic differences between them. Another interesting point could be in comparing valency patterns of attributive and predicative uses of adjectives based on FrameBank data. The assessment of the new role inventory in the existing SRL modules will also be helpful for both the dictionary tasks and for the development of automatic semantic analysis for Russian.

References

1. *Anisimovich K., Druzhkin K., Minlos F., Petrova M., Selegey V., Zuev K.* (2012). Syntactic and Semantic parser based on ABBYY Compreno linguistic technologies, in Proceedings of the International Conference “Dialogue”, Vol. 11–2, pp. 91–103.
2. *Apresjan Ju. D.* (1995), Selected papers, Vol. 1, Lexical Semantics [Izbrannye trudy, tom I. Leksicheskaja semantika], Jazyki Russkoj Kul’tury, Vostochnaja Literatura, Moscow.
3. *Apresjan Ju. D.* (ed.) (2004), New Russian Explanatory Synonym Dictionary [Novyj ob”iasnitel’nyj slovar’ sinonimov russkogo jazyka], Jazyki slavjanskoj kul’tury, Moscow.
4. *Apresjan Ju. D., Boguslavskij I. M., Iomdin L. L., Sannikov V. Z.* (2010), Theoretical issues of Russian syntax: the interrelation between grammar and vocabulary [Teoreticheskie problemy russkogo sintaksisa: vzaimodejstvie grammatiki i slovarja], Jazyki slavjanskih kul’tur, Moscow.
5. *Bonial C., Bonn J., Conger K., Hwang J., Palmer M., Reese N.* (2015). English PropBank Annotation Guidelines. Center for Computational Language and Education Research, Institute of Cognitive Science, University of Colorado at Boulder.
6. *Bulygina T. V., Shmelev A. D.* (1997), Language conceptualization of the world (based on Russian grammar) [Jazykovaja konceptualizacija mira (na materiale russkoj grammatiki)], Jazyki Russkoj Kul’tury, Moscow.

7. *Callison-Burch C., Fordyce C., Koehn Ph., Monz C., Schroeder J.* (2007), (Meta-) Evaluation of machine translation, in Proceedings of WMT, pp. 136–158.
8. *Dowty D. R.* (1991), Thematic proto roles and argument selection, Language, Vol. 67, pp. 547–619.
9. *Dras M.* (2015), Evaluating human pairwise preference judgements, Computational Linguistics, Vol. 41 (2), pp. 309–317.
10. *Faruqui M., Dyer C.* (2014), Community Evaluation and Exchange of Word Vectors at wordvectors.org, in Proceedings of System Demonstrations, ACL 2014, available at: http://www.cs.cmu.edu/~mfaruqui/papers/acl14-vecdemo.pdf
11. *Fillmore Ch. J.* (1968), The Case for Case, in Bach E. and Harms (Ed.), Universals in Linguistic Theory. New York, pp. 1–88.
12. *FrameNet*. An online resource, available at: http://framenet.icsi.berkeley.edu
13. *Karpova O. S., Reznikova T. I., Arkhangel'skij T. A., Kjuseva M. V., Rakhilina E. V., Ryzhova D. A., Tagabileva M. G.* (2010), A database of polysemous qualitative adjectives and adverbs in Russian [Baza dannyh po mnogoznachnym kachestvennym prilagatel'nym I narechijam russkogo jazyka], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'juternaja lingvistika i intelleltual'nye tehnologii: po materialam ezhegodnoj Mezhdunarodnoj konferentsii "Dialog"], Moscow, pp. 163–168.
14. *Kashkin E. V., Lyashevskaya O. N.* (2013), Semantic roles and construction net in Russian FrameBank [Semanticheskie roli i set' konstrukcij v sisteme FrameBank], Computational linguistics and intellectual technologies. Proceedings of International Conference "Dialog", Vol. 12–1, pp. 297–311.
15. *Kustova G. I.* (2007), Russian adjectives: forms, constructions, semantics [Russkie prilagatel'nye: formy, konstruktsii, semantika], in Nauchnye trudy Moskovskogo pedagogicheskogo gosudarstvennogo universiteta. Filologicheskie nauki [Proceedings of Moscow pedadogical statue university. Philology], Prometej Publ., Moscow, pp. 85–97.
16. *Kustova G. I.* (2009), Elektronnyj semanticheskij slovar' glagol'nyh prilagatel'nyh: struktura i tipy informacii [Electronic semantic dictionary of deverbal adjectives: structure and types of information], Computational linguistics and intellectual technologies. Proceedings of International Conference "Dialog", Vol. 8, pp. 271–277.
17. *Kuznetsov I. O.* (2015), Semantic Role Labeling for Russian Language Based on Russian FrameBank, in M. Yu. Khachay, N. Konstantinova, A. Panchenko, D. I. Ignatov, V. G. Labunets (eds.), Analysis of Images, Social Networks and Texts. Fourth International Conference, AIST 2015, Yekaterinburg, Russia, April 9–11, 2015, Revised Selected Papers. Communications in Computer and Information Science, Vol. 542, Springer, pp. 333–338.
18. *Liu B.* (2012), Sentiment analysis and opinion mining, Synthesis lectures on human language technologies, Vol. 5(1), pp. 1–167.
19. *Lyashevskaya O.* (2010), Bank of Russian Constructions and Valencies, Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, pp. 1802–1805.

20. *Lyashevskaya O. N., Kashkin E. V.* (2015a), FrameBank: a database of Russian lexical constructions, in M. Yu. Khachay, N. Konstantinova, A. Panchenko, D. I. Ignatov, V. G. Labunets (eds.), Analysis of Images, Social Networks and Texts. Fourth International Conference, AIST 2015, Yekaterinburg, Russia, April 9–11, 2015, Revised Selected Papers. Communications in Computer and Information Science, Vol. 542, Springer, pp. 350–360.
21. *Lyashevskaya O. N., Kashkin E. V.* (2015b), Types of information about lexical constructions in Russian FrameBank [Tipy informatsii o leksicheskih konstruktsijah v sisteme FrameBank], Trudy Instituta russkogo jazyka im. V. V. Vinogradova [Proceedings of the V. V. Vinogradov Russian Language Institute], Vol. 6, pp. 464–555.
22. *Màrquez L., Carreras X., Litkowski K. C., Stevenson S.* (2008), Semantic role labeling: an introduction to the special issue, Computational Linguistics, Vol. 34–2, pp. 145–159.
23. *Paducheva E. V.* (2004), Dynamic patterns in lexical semantics [Dinamicheskie modeli v semantike leksiki], Jazyki slavjanskoj kul'tury, Moscow.
24. *Palmer M., Gildea D., Kingsbury P.* (2005), The Proposition Bank: An annotated corpus of semantic roles, Computational Linguistics, Vol. 31–1, pp. 71–106. https://github.com/propbank/propbank-documentation/raw/master/annotation-guidelines/Propbank-Annotation-Guidelines.pdf
25. *Palmer M. S., Wu Sh., Titov I.* (2013), Semantic Role Labeling Tutorial. NAACL 2013 tutorials. Available at: http://naacl2013.naacl.org/Documents/semantic-role-labeling-part-1-naacl-2013-tutorial.pdf, http://naacl2013.naacl.org/Documents/semantic-role-labeling-part-2-naacl-2013-tutorial.pdf, http://naacl2013.naacl.org/Documents/semantic-role-labeling-part-3-naacl-2013-tutorial.pdf
26. *Rakhilina E. V., Reznikova T. I., Karpova O. S.* (2010), Semantic shifts in attributive constructions: metaphor, metonymy, and rebranding [Semanticheskie perehody v atribultivnyh konstrukcijah: metafora, metonimija i rebrending], in Linguistics of Constructions [Lingvistika konstrukcij], Azbukovnik, Moscow, pp. 398–455.
27. *Shelmanov A. O., Smirnov I. V.* (2014), Methods for semantic role labeling of Russian texts, Computational linguistics and intellectual technologies. Proceedings of International Conference "Dialog", Vol. 13, pp. 607–619.
28. *Vol'f E. M.* (1978), Grammar and Semantics of Adjectives. Based on Ibero-Roman Languages. [Grammatika i semantika prilagatel'nykh. Na materiale ibero-romanskikh jazykov], Nauka, Moscow.

FORMAL MODELING OF CASE VARIATION: A PARAMETRIC APPROACH[1]

Lyutikova E. A. (lyutikova2008@gmail.com)

MSU/IMLR MPSU, Moscow, Russia

This paper aims at evaluating formal theories of case assignment with respect to their applicability to modeling of case variation. Crosslinguistically, differential case marking exhibit significant variation in many parameters, including licensing factors of case variation, correlation of case with linear position, and feeding of predicative or possessive agreement. In this paper, I consider the two most elaborated formal theories of case—the minimalist syntactic case theory and the configurational case theory—and explore their expressive power in modeling various types of differential case marking. I show that none of the theories is superior to the other—rather, each of them naturally accommodates a specific type of case variation but is unsuitable to express the other types. The minimalist syntactic case theory is more flexible in that it is compatible with additional mechanisms deriving the morphologically observable case variation, and more restrictive in that it predicts the one-to-one correspondence between case assignment and agreement. The prime advantage of the configurational theory is that it can represent directly the non-local dependencies between case-marking of different arguments.

Key words: formal language models, typology, case, agreement, differential case marking

ФОРМАЛЬНОЕ МОДЕЛИРОВАНИЕ ПАДЕЖНОГО ВАРЬИРОВАНИЯ: ПАРАМЕТРИЧЕСКИЙ ПОДХОД

Лютикова Е. А. (lyutikova2008@gmail.com)

МГУ/ИСЛИ МПГУ, Москва, Россия

Ключевые слова: формальные модели языка, типология, падеж, согласование, дифференцированное падежное маркирование

[1] The research has been supported by Russian Scientific Foundation (project № 14-18-03270 “Word order typology, communicative-syntactic interface and information structure in the world’s languages”).

1. Introduction

Case marking is one of the most reliable linguistic "tags" that signal the relation of its bearer to the syntactic context and/or its thematic relation to the governing element. The correct recognizing of the argument structure of predicates is crucial for understanding the core meaning of the sentence in both human and automatic language processing. Accordingly, using case morphology as a marker of a specific grammatical relation of the nominal to the predicate/sentence or as an encoding of a specific thematic role is a universal mechanism of syntactic parsing and semantic analysis.

Indeed, research in both experimental and computational linguistics shows the role of case marking in the parsing process. A facilitation effect in sentence parsing due to case marking has been reported for many languages with rich case morphology, such as Japanese (Mazuka, Lust 1990; Yamashita 1997), Finnish (Hujanen 1997) or Basque (Santesteban et al. 2015). Similarly, syntactic parsing of morphologically rich languages can rely heavily on morphological dependencies (Tsarfaty et al. 2013). Thus, using of morphological marking of the highly inflectional language such as Czech in adjusting a statistical parsing model showed a 7% improvement in dependency accuracy (Collins et al. 1999); see also (Øvrelid, Nivre 2007; Øvrelid 2008) on argument differentiation based on different grammatical features in data-driven dependency parsers.

The one-to-one correspondence between morphological case marking and syntactic relation is broken in at least two directions. The first challenge is the case syncretism that results in underdetermination of the morphological tagging of a nominal and makes it much harder for a statistical system to learn the morphological marking patterns of a language (Seeker, Kuhn 2013). The second challenge is that the regular case marking of the given grammatical function can be biased by the case variation process licensed by some language-specific grammatical factor.

Although differential case marking (DCM) is highly widespread, its impact on language processing seems to be underestimated (but see (Butt, King 2001; Butt, King, Varghese 2004)). Yet, modeling of DCM can neutralize the disturbing effect of case variation on the association of a nominal with its grammatical or thematic role; moreover, factors licensing differential case marking can contribute to the syntactic and semantic analysis.

In this paper, I investigate modeling of case variation within the two competing formal theories of case—the minimalist syntactic case theory and the configurational case theory. I use the parameters distinguishing cross-linguistically attested patterns of DCM in order to identify the characteristic "profile" of case variation that can be easily accommodated within each theory, as well as the blind-spots of each theory, that is, linguistic phenomena that cannot be expressed within it.

The paper is organized as follows. Section 2 provides a typological overview of differential case marking and introduces parameters that characterize case variation. In section 3, the two formal theories of case assignment are briefly described. Section 4 discusses their potential in modeling different aspects of case variation. Conclusions are drawn in section 5.

2. Differential case marking: positions and parameters

Case variation is a widespread linguistic phenomenon. The most deeply studied differential object marking (DOM) received much attention in the literature at least since (Bossong 1985); the keynote papers summarizing the functional and typological perspective of DOM are Aissen 2003; de Swart 2007; von Heusinger, Klein and de Swart 2008; de Hoop, Malchukov 2007; Malchukov, de Swart 2008. In (1), a definiteness-conditioned DOM in Hebrew is exemplified.

(1) a. *Dan* *kara* **(**et**)* ***ha-itonim***.
Dan read OM DEF-newspapers
'Dan read the newspapers.'

b. *Dan* *kara* *(***et**)* ***itonim***.
Dan read (*OM) newspapers
'Dan read newspapers.' (Danon 2006: (1))

Differential subject marking (DSM) was recognized as a unified phenomenon only recently (de Hoop, de Swart (eds.) 2009); previous approaches to DSM focused mostly on split ergativity—cases when ergative alignment is absent in some transitive clauses (Silverstein 1976; deLancey 1981; Dixon 1994) and on active/semantic alignment or split intransitivity, whereby the sole argument of an intransitive verb does not receive a uniform encoding (Mithun 1991; Donohue, Wichman (eds.) 2008). DSM is often accompanied by DOM: thus, in aspectually determined split ergativity in Georgian ergative/nominative encoding of the subject vary in parallel with nominative/dative encoding of the object, as in (2).

(2) a. Aorist (Series II), ergative subject, nominative object
nino-m *gia-s* ***surateb-i*** *ačvena.*
Nino-ERG Gia-DAT picture-NOM show.AOR
'Nino showed the picture to Gia.'

b. Present (Series I), nominative subject, dative object
nino *gia-s* ***surateb-s*** *ačvenebs.*
Nino.NOM Gia-DAT picture-DAT show.PRS
'Nino is showing the picture to Gia.'

A huge amount of data concerns differential case marking of arguments in embedded structures such as nominalizations or causative constructions (Comrie 1976; Givón 1990; Alexiadou 2001). In nominalizations, case variation usually involves "clausal" and "nominal" case marking of the subject (Szabolcsi 1983; Abney 1987), and this is how possessor case marking is included into the perspective of the differential case marking of clausal arguments. Example (3) shows specificity-induced DSM in Turkish nominalized clause where "clausal" nominative case marking varies with "nominal" genitive case marking (von Heusinger, Kornfilt 2005: (16)).

(3) a. *Yol-dan* ***bir*** ***araba*** *geç-tiğ-in-ı* *gör-dü-m.*
road-ABL one car pass-NML-3SG-ACC see-PST-1SG
'I saw that a car [non-specific] went by on the road.'

b. *Yol-dan* ***bir*** ***araba-nın*** *geç-tiğ-in-ı* *gör-dü-m.*
road-ABL one car-GEN pass-NML-3SG-ACC see-PST-1SG
'I saw that a car [specific] went by on the road.'

The morphosyntactic encoding of the possessor, in its turn, exhibit variation concerning semantic type of possessive relation (alienable/inalienable possession, Haiman 1983), structural position of the possessor (internal/external possessor, Shibatani 1994; Payne, Barshi (eds.) 1999) and its referential characteristics (Pereltsvaig, Lyutikova 2014; Testelets 2014).

Thus, DCM occurs in various syntactic positions; it should be noted, however, that these positions—subject, object, possessor—have much in common: they correspond to the basic grammatical functions of the noun phrases that tend to receive a uniform linguistic encoding—special flagging with grammatical cases (nominative, accusative, ergative, genitive) and priority indexing in the predicate (predicative and possessive agreement). This prototypical encoding is biased by licensing factors of DCM to the effect that the nominal receives a special case marking (or "loses" case marking), and this can influence its ability to control agreement. This general scheme of case variation, as well as the regular interrelations of DOM, DSM and possessor marking mentioned above allow us to analyze DCM within a single parametric system, extending and adjusting the parametric description of DOM proposed in (von Heusinger, Klein and de Swart 2008).

At least the following parameters of DCM required by empirical generalizations and relevant for formal models of case assignment can be distinguished:

- locality;
- semantic motivation;
- positional differences;
- correlation with agreement.

Degree of **locality** characterizes the localization of the licensing factor of DCM. Local DCM is conditioned by the properties of the argument itself—e.g. its formal feature (noun/pronoun, locutor/non-locutor), syntactic category (DP/NP), animacy, definiteness, referentiality, topicality etc. Non-local DCM comes in two varieties: predicate-determined and coargument-determined DCM. Predicate-determined DCM is based on properties of the predicate (possibly in combination with the properties of the argument). Thus, DOM in Finnic languages involves quantization of the internal argument and perfectivity/imperfectivity of the VP (4):

(4) a. *Ta* *ehitas* ***silla*** *(kahe* *aasta-ga).*
he build.PST bridge.GEN (two.GEN year.OBL-COMIT)
'He built a bridge (in two years).'

b. *Ta* *ehitas* ***silda*** *(kaks* *aastat).*
he build.PST bridge.PART (two.PART year.PART)
'He was building a bridge (for two years).'

The aspectual, temporal and modal characteristics of the predicate can also influence DSM. In Hindi, ergative encoding of the external argument is attested in perfective clauses, whereas imperfective clauses license a nominative subject (5). Other properties of the predicate that can license DSM of the external argument are volitionality and agentivity; they often correlate with animacy of the external argument.

(5) a. ***laṛke-ne*** *kitāb* *xarīdī.*
boy-ERG book.NOM buy.PF
'The boy bought a book.'

b. ***laṛkā*** *kitāb* *xarīdtā* *hai.*
boy.NOM book.NOM buy.IPF AUX.PRS
'The boy is buying a book.'

Coargument-determined DCM takes place if an argument gets a case marking depending on the characteristics of its coargument. In (6) from Awtuw (Sepik-Ramu) the direct object receives accusative marking only if it overranks the subject in animacy (de Hoop, Malchukov 2007); so DOM in Awtuw is determined hierarchically and motivated by disambiguation (but see Arkadiev 2008 for the critisism of disambiguation approach based on DCM in two-term case systems).

(6) a. *Tey* ***tale-re*** *yaw* *d-æl-i.*
3F.SG woman-ACC pig F.AGT-bite-PAT
'The pig bit the woman.'

b. *Tey* ***tale*** *yaw* *d-æl-i.*
3F.SG woman pig F.AGT-bite-PAT
'The woman bit the pig.'

DSM can be conditioned by the encoding of the internal argument: in Bagwalal (Andic/Dagestanian) example (7), the choice of the ergative or nominative case of the external argument depends on the case marking of the internal argument (which reflects its thematic role).

(7) a. ***anwar*** */ *anwar-i-r* ***ila-ɬa*** *w-alli.*
Anvar.NOM / *Anvar-OBL-ERG mother-OBL.SUP.LAT M-shout.PST
'Anvar addressed to his mother.'

b. ***anwar-i-r*** */ *anwar* ***ila*** *j-alli.*
Anvar-OBL-ERG / *Anvar.NOM mother.NOM F-shout.PST
'Anvar called his mother.'

The parameter of **semantic motivation** distinguishes between the semantically motivated vs purely configurational case variation. Let us consider the encoding of the causee in the causative construction. In Balkar (Turkic/Altaic), causee encoding observes Comrie's (1976) Paradigm Case rule and is determined exclusively by transitivity of the input verb. Regardless of the thematic role of the causee, his/her control over the performed action and semantics of the causative construction, the causee receives accusative with causatives of intransitives and dative with causatives of transitives (Lyutikova et al. 2006). In Hungarian, however, the case of the causee depends on its agentivity: in causer-controlled structures, the causee receives accusative, but in causee-controlled structures, it is marked with instrumental.

(8) a. *Az* *orvos* *pisiltette* ***a*** ***gyereket***.
DET doctor.NOM pee.CAUS.3SG DET child.ACC
'The doctor made the child pee.'

b. *Az* *orvos* *pisiltetett* ***a*** ***gyerekkel***.
DET doctor.NOM pee.CAUS.3SG DET child.INSTR
'The doctor had the peeing done by the child.' (Ackerman 1994:537)

Positional distribution is often observed with DCM. In Sakha, accusative and unmarked direct objects occupy different positions with respect to the indirect object or VP-level adverbials ((9), Baker, Vinokurova 2010).

(9) a. *Masha* [$_{VP}$ ***türgennik*** ***salamaat-(*y)*** *sie-te*].
Masha quickly porridge-(*ACC) eat-PST.3SG

b. *Masha* ***salamaat-*(y)*** [$_{VP}$ ***türgennik*** *sie-te*].
Masha porridge-*(ACC) quickly eat-PST.3SG

(a=b) 'Masha ate porridge quickly.'

Similarly, in Tatar, genitive case-marked referential possessor precedes adjectives, whereas unmarked non-referential possessor follows them (Pereltsvaig, Lyutikova 2014; Lyutikova, Pereltsvaig 2015). However, case variation may occur in the same structural position. This becomes clear if we compare Tatar example (10) with Sakha example (9). In Tatar, accusative direct object can occupy the same (preverbal) position as the unmarked direct object.

(10) *Bajras* *kat-kat* ***xat-(nı)*** *ukı-dı.*
Bayras again-again letter-(ACC) read-PST
'Bayras read the/a letter again and again.'

The **correlation** of DCM **with agreement** is evident when one (or even both) of the cases that can be assigned to a noun phrase licenses predicative or possessive agreement. In Tatar example (11), the subject of the relative clause exhibit case

variation between nominative and genitive. Genitive case marking enforces possessive agreement on the head noun.

(11) a. ***Marat*** *Kazan-nan* *al-ıp* *kajt-kan* ***kitap*** *bik* *kızık.*
Marat Kazan-ABL take-CONV return-PF book very interesting

b. ***Marat-nıŋ*** *Kazan-nan* *al-ıp* *kajt-kan* ***kitab-ı*** *bik* *kızık.*
Marat-GEN Kazan-ABL take-CONV return-PF book-3 very interesting

(a=b) 'The book that Marat brought from Kazan is very interesting.'

On the other hand, agreement may persist irrespectively of case marking. This is what happens in Amharic (Ethiopian/Semitic, Baker 2012) where primary object can be marked with accusative or embedded under the prepositional phrase. Object agreement is optional with accusative primary object (12a) and possible with prepositional primary object (12b). Thus, case variation in primary objects does not influence agreement options.

(12) a. *Ləmma* ***wɨʃʃa-w-ɨn*** *j-aj-al* || *j-aj-**əw**-al.*
Lemma dog-DEF-ACC 3M.SU-see-AUX.3M.SU 3M.SU-see-3M.OBJ-AUX.3M.SU
'Lemma sees the dog.'

b. ***L-Aster*** *lɨʤ-u-n* *assajj-*əhw-***at***.
DAT-Aster baby-DEF-ACC show-1SG.SU-3F.OBJ
'I showed the baby to Aster.'

In section 4, I discuss the possible interpretation of the parameters outlined above within the syntactic models of case, but first, a brief characterization of the two formal theories of case assignment is due.

3. Formal theories of case assignment

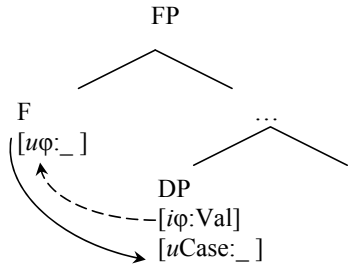
In the formal syntactic literature, two major approaches to case assignment can be found. The first approach, which is mainly associated with Noam Chomsky's work, considers case as a syntactic phenomenon that licenses NPs; the second approach, put forward in the work by Alec Marantz, treats case as a postsyntactic, purely morphological phenomenon.

In the modern minimalist syntactic approach (Chomsky 2000, 2001) Case is an unvalued uninterpretable feature of a noun phrase that has to be valued in the course of the derivation. In the Chomsky-style model, Case is assigned to a noun phrase under AGREE relation with a dedicated case-assigning head.

In (13), the head F which has unvalued φ-features [*u*φ:_] acts as a goal and seeks for an active bearer of valued φ-features [*i*φ:Val] in its c-command domain. The DP with an unvalued Case and valued φ-features suits as a goal. The AGREE relation

between F and DP is established, and F values its φ-features copying the values accessible on the DP. The φ-complete F can, in its turn, value the case feature of the goal. Additionally, the second-order feature [EPP] on [*u*φ:_] may attract the goal DP into the projection of FP.

(13)



Two kinds of case-assigning heads can be distinguished: lexical heads, that assign case to their own arguments exclusively, and functional heads, that assign structural case to the nearest goal DP available in their c-command domain. Lexical heads assign the lexical (or inherent) case at the very moment of merging with their arguments, discharging their theta-roles; thus lexical case is syntactically local and theta-related.

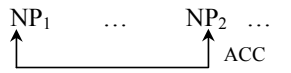
The characteristic properties of a structural case are: (i) its independence from a theta-role; (ii) its somewhat non-local nature and (iii) the non-obligatoriness of its realization. In view of these properties, three major structural cases are usually recognized: nominative, assigned by the finite predicative head T; accusative, assigned by the transitive light verb head *v*; and genitive, assigned by the (possessive) determiner head D.

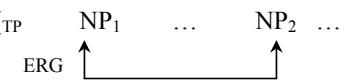
The competing configurational approach dates back to the seminal paper by Alec Marantz (Marantz 1991). The basic idea behind this approach is that (morphological) case assignment can be dependent not only on the presence of governing heads, but also on the presence of other noun phrases ("case competitors") in the same syntactic domain.

The category of case is organized hierarchically. Marantz distinguishes four distinct kinds of case, forming a disjunctive Case realization hierarchy (14). This hierarchy determines the order in which the different kinds of case shall be assigned.

(14) Case realization disjunctive hierarchy:
- lexically governed case
- "dependent" case (accusative and ergative)
- unmarked case (environment-sensitive)
- default case

First, the most specific lexically-governed case is assigned. Next, the rule of dependent case assignment applies. The dependent case rule requires a configuration where there are at least two caseless NPs in the clausal domain. If this requirement is met, these noun phrases enter into case-competition. In accusative languages, the lower NP is marked with the "dependent" accusative case (15a), and in ergative languages, the higher NP is marked with the "dependent" ergative case (15b).

(15) a. 

b. 

Then, the rule of the unmarked case applies that marks any still case-less NP in a given syntactic domain with the dedicated case. Finally, if neither of the previous rules applied to an NP, it receives the default case. It is important that the universal availability of the default case realization in Marantz's system means that case assignment is set apart from licensing: case as a purely morphological phenomenon only interprets the syntactic structure, but does not filter it out.

Though some adherents of the approach pursue the postsyntactic analysis of case issues (McFadden 2004; Bobaljik 2008), attempts have been made to incorporate the appealing idea of the "dependent" case assignment into the narrow syntax. Thus, in (Bittner, Hale 1996; Baker 2009, 2012, 2015; Preminger 2011, 2014; Kornfilt, Preminger 2015; Levin, Preminger 2015) the authors explore various paths of implementing configurational case assignment within the syntactic model of case. The basic innovations in the Marantz-style system include a more elaborate definition of case competition domains. Within the clause, more domains for case competition have been distinguished, e.g. VP and TP, which allowed to subsume dative case assigned in ditransitive constructions under a similar analysis.

Although Marantz-style case assignment is construed as independent from agreement of lexical or functional heads (i.e. AGREE operation), the morphological case marking can in principle feed the agreement process (Bobaljik 2008; Preminger 2011, 2014). Thus, J. Bobaljik reinterprets E. Moravcsik's (1974, 1978) hierarchy as the hierarchy of accessibility of case-marked DPs as controllers of agreement (16). Agreement is said to be case-discriminating, in the sense that only those DPs that bear a specific case are visible as goals for a probe looking for a source of valued φ-features.

(16) unmarked case ≫dependent case ≫ lexical case

In the series of papers by J. Kornfilt (Kornfilt 2013; Kornfilt, Preminger 2015) a further refinement of the configurational analysis is put forward: "nominative" and "genitive" are considered as descriptive labels for caseless NPs. In her system, agreement targets caseless NPs exclusively.

In the next section, I address the question of how case variation can be implemented within the proposed systems.

4. Evaluating theories

Let us start with discussing explanatory mechanisms available for the two theories. Lexical case assignment seems to be treated similarly in both of them, so the difference is in explaining DCM involving structural cases.

It seems that the configurational case assignment model has only one explanatory mechanism based on the positional alternative. If a noun phrase exhibit differential structural case marking, it may belong to different domains of case assignment, or its domain of case assignment may contain or lack a potential case competitor. This is how various factors licensing structural case variation are to be modeled.

Chomsky-style case assignment is more flexible in that it allows various mechanisms of modeling factors triggering case variation. External factors like telicity, perfectivity or polarity are naturally conceived as (features of) functional heads that assign a structural case to DPs. The impact of factors internal to the noun phrase can be analyzed as a split morphological realization of the same syntactic case. Besides, as case assignment obeys at least phase-level locality and relativized minimality, the positional alternative is also an option for the Chomsky-style model.

Now we can proceed to the parameters of DCM.

The **locality** parameter distinguishes the local DCM and the two types of the nonlocal DCM: predicate-determined and coargument-determined. The local DCM is easily implemented in the Chomsky-style syntactic case theory, e.g. as a morphologically conditioned realization of the syntactic case, as a parametrization of the case filter allowing structurally deficient nominals to remain caseless (Lyutikova 2014) or as positionally distributed case options available for noun phrases with different features. Configurational case theory can only rely on the positional alternative. Predicate-determined DCM is best conceived as the interdependence between a predicate with a specific feature and an argument selected by it. This is exactly what the "case assignment by functional heads" theory proposes. For the configurational case theory, modeling of predicate-determined DCM has to be mediated by the reorganization of structural configurations, whereby the licensing factor (e.g. imperfective aspect) is associated with more case domains than its counterpart (i.e. perfective aspect). Although sometimes it is indeed the case (thus, biabsolutive constructions in Basque or Tsez have been claimed to be biclausal), extending this mechanism to all cases of predicate-determined DCM is not an attractive decision. On the contrary, coargument-determined DCM is best subsumed under the configurational model. In example (7) above DSM is easily modeled as depending on the presence or absence of the case competitor—the internal argument without lexical case. For a Chomsky-style model, the analysis of (7) involves ergative-assigning and nominative-assigning functional heads, and their choice should somehow depend on the case assigned to the internal argument.

In both models, **semantic motivation** can be easily implemented if it yields the lexical case assignment (cf. Genitive of negation in Russian). If the DCM involves two structural cases, the explanatory proficiency of the two theories differs. In the minimalist theory the regular semantically motivated case, such as instrumental in (8b), can be interpreted as a theta-related case assigned to an argument by a functional head (this is E. Woolford's (2006) analysis of ergative and dative). In the configurational case theory, however, all non-lexical cases (such as ergative, dative, or instrumental) are semantically as empty as the accusative or nominative.

Positional distribution accompanying the DCM is in the very heart of the configurational model. If case variation occurs in the same syntactic position and involves structural cases, as is the case in Tatar DOM (10), the configurational model

is handicapped. As for the Chomsky-style model, it can deal with both positionally dependent and independent DCM.

The interdependence of case assignment and **agreement** is an attribute of the minimalist case theory. It presupposes one-to-one correspondence between agreement and case, because valuing the φ-features of the functional head enables it to assign case to the goal DP. If no overt agreement is associated with case marking, one can hypothesize that the AGREE operation still takes place but is not realized morphologically. It is much more difficult to explain the optionality of agreement (12a) and the immunity of agreement to case variation (12b).

The configurational model accounts easily for various kinds of splits between agreement and case assignment. If a strong correlation between case and agreement exists (e.g. only nominative subjects control predicative agreement, and predicative agreement is obligatory if there is a nominative subject), the mechanism of case discrimination can be exploited. The weak point of this mechanism is that it predicts the possibility of the multiple agreement of different heads with the same DP. If this option is undesirable on empirical grounds, the configurational model fails to exclude it.

Evaluation of case theories in representing various types of DCM is summarized in Table 1.

Table 1. Parameters of case variation in the formal theories of case assignment

Parameter	Minimalist case theory	Configurational case theory
Locality	local DCM predicate-determined DCM	local DCM coargument-determined DCM
Semantic motivation	easily representable	non-representable outside of lexical government
Positional distribution	non-obligatory	obligatory
Correlation with agreement	strong correlation	various splits between case and agreement

5. Conclusions

In this paper, I considered the two most elaborated formal theories of case—the minimalist syntactic case theory and the configurational case theory—and explored their expressive power in modeling various types of differential case marking. I showed that none of the theories is superior to the other—rather, each of them has its own strengths and weaknesses in modeling different types of case variation. However, this conclusion should not disappoint us. It seems that the mere existence of various patterns of DCM calls for the elaboration of various models of case assignment—at least until the uniform theory of case, flexible enough to account equally well for all attested types of DCM, and restrictive enough to exclude unattested types,

is proposed. Meanwhile, we should be aware of different theories, their potential and their limitations, in order to choose the right model for the empirical data.

Addressing applicability of the study to NLP, I shall emphasize that linguistic rules possibly employed in argument structure retrieval are not required to constitute a uniform theoretical system. Thus, we can adjust the specific mechanisms of the two models discussed in this papers to specific DCM phenomena of a specific language, to the effect that, for example, in Balkar the identification of the causee can be determined configurationally, but the identification of the possessor can be based on the presence of an agreeing nominal head. Interestingly, such a hybrid theoretical system aiming at more natural relations of facts and theories has been proposed recently for Sakha (Baker, Vinokurova 2010). It seems that exactly this sort of models is in demand in computational linguistics.

Abbreviations

1—1st person; 3—3rd person; ABL—ablative; ACC—accusative; AGT—agent; AOR—aorist; AUX—auxiliary; CAUS—causative; COMIT—comitative; CONV—converb; D—determiner (syntactic category); DAT—dative; DCM—differential case marking; DEF—definite; DET—determiner (lexical item); DOM—differential object marking; DP—determiner phrase; DSM—differential subject marking; EPP—extended projection principle; ERG—ergative; F—feminine; GEN—genitive; INSTR—instrumental; IPF—imperfective; LAT—lative; M—masculine; NML—nominalization; NOM—nominative; NP—noun phrase; OBJ—object agreement; OBL—oblique stem; OM—object marker; PART—partitive; PAT—patient; PF—perfective; PL—plural; PRS—present; PST—past; SG—singular; SU—subject agreement; SUP—super (localization); TP—tense phrase (clause); VP—verbal phrase; φ-features—person, number, nominal class features.

References

1. *Abney, Steven* (1987). The English noun phrase in its sentential aspect. PhD thesis, MIT, Cambridge (MA).
2. *Ackerman, Farrell* (1994). Entailments of predicates and the encoding of causees. Linguistic Inquiry Vol. 25 (3), pp. 535–547.
3. *Aissen, Judith* (2003). Differential Object Marking: Iconicity vs. Economy. Natural Language and Linguistic Theory, Vol. 21, pp. 435–483.
4. *Alexiadou, Artemis* (2001). Functional structure in nominals: Nominalization and ergativity. Amsterdam, Philadelphia: John Benjamins.
5. *Arkadiev, Peter* (2008). Poor (two-term) case systems: limits of neutralization. In: A. Malchukov and A. Spencer (eds), Handbook of Case. Oxford University Press, pp. 686–699.
6. *Baker, Mark C.* (2009). Is head movement still needed for noun incorporation? The case of Mapudungun. Lingua, Vol. 119 (2), pp. 148–165.

7. *Baker, Mark C.* (2012). On the relationship of object agreement and accusative case: Evidence from Amharic. Linguistic Inquiry, Vol. 43, pp. 255–274.
8. *Baker, Mark C.* (2015). Case. Its principles and its parameters. (Cambridge studies in linguistics 146) Cambridge: CUP.
9. *Baker, Mark C., and Nadya Vinokurova* (2010). Two modalities of Case assignment: Case in Sakha. Natural Language and Linguistic Theory, Vol. 28, pp. 593–642.
10. *Bittner, Maria, and Kenneth Hale* (1996). The Structural Determination of Case and Agreement. Linguistic Inquiry, Vol. 27, pp. 1–68.
11. *Bobaljik, Jonathan* (2008). Where's phi? Agreement as a post-syntactic operation. In D. Harbour et al. (eds.). Phi Theory, Oxford: Oxford University Press, pp. 295–328.
12. *Bossong, Georg* (1985). Empirische Universalienforschung. Differentielle Objektmarkierung in der neuiranischen Sprachen. Tübingen: Narr.
13. *Butt, Miriam, and Tracy Holloway King* (2001). Non-nominative subjects in Urdu: A computational analysis. In Proceedings of the International Symposium on Non-nominative Subjects, Tokyo, ILCAA, pp. 525–548,
14. *Butt, Miriam, Tracy King, and Anila Varghese* (2004). A computational treatment of differential case marking in Malayalam. International Conference on Natural Language Processing, 2004 December 19–22, Hyderabad, India.
15. *Chomsky, Noam* (2000). Minimalist Inquiries: The Framework. In R. Martin, D. Michaels, and J. Uriagereka (eds.). Step by Step. Essays on Minimalist Syntax in Honor of Howard Lasnik. Cambridge, MA: MIT Press, pp. 89–155.
16. *Chomsky, Noam* (2001). Derivation by Phase. In M. Kenstowicz (ed.). Ken Hale: A Life in Language. Cambridge, MA: MIT Press, pp. 1–52.
17. *Collins, Michael, Jan Hajič, Lance Ramshaw, and Christoph Tillmann* (1999). A statistical parser for Czech. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, MD, pp. 505–512.
18. *Comrie, Bernard* (1976). The syntax of causative constructions: cross-language similarities and divergences. In Masayoshi Shibatani, ed.: The Grammar of Causative Constructions (Syntax and Semantics 6), New York: Academic Press, pp. 261–312.
19. *Danon, Gabi* (2006). Caseless Nominals and the Projection of DP. In: Natural Language and Linguistic Theory, Vol. 24, pp. 977–1008.
20. *de Hoop, Helen, and Peter de Swart* (eds.) (2009). Differential Subject Marking. Dordrecht: Springer.
21. *de Hoop, Helen, and Andrej L. Malchukov* (2007). On fluid differential case marking: A bidirectional OT approach. Lingua, Vol. 117, pp. 1636–1656.
22. *de Swart, Peter* (2007). Cross-linguistic Variation in Object Marking. Doctoral dissertation, University of Nijmegen.
23. *DeLancey, Scott* (1981). An interpretation of split ergativity and related patterns. Language, Vol. 57 (3), pp. 626–57.
24. *Dixon, R. M. W.* (2004). Ergativity. Cambridge: Cambridge University Press.
25. *Donohue, Mark, and Søren Wichman* (eds.) (2008). The Typology of Semantic Alignment. Oxford: Oxford University Press.
26. *Givón, Talmy* (1990). Syntax: A Functional-Typological Introduction. Vol. II. Philadelphia: John Benjamins.

27. *Haiman, John* (1983). Iconic and economic motivation. Language, Vol. 59, pp. 781–819.
28. *Hujanen, Jukka* (1997). Effects of Case Marking and Word Order on Sentence Parsing in Finnish: An Eye Fixation Analysis. The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, Vol. 50 (4), pp. 841–858.
29. *Kornfilt, Jacklin* (2013). Nominative as no case at all: An argument from raising-to-accusative in Sakha. Paper presented at WAFL9 workshop, August 23–25, 2013, Cornell University, Ithaca, NY.
30. *Kornfilt, Jacklin, and Omer Preminger* (2015). Nominative as no case at all: an argument from raising-to-accusative in Sakha. In Proceedings of the 9th Workshop on Altaic Formal Linguistics (WAFL9) / Ed. by Andrew Joseph and Esra Predolac (MIT Working Papers in Linguistics #76). Cambridge (Mass), MIT Press, pp. 109–120.
31. *Levin, Theodor, and Omer Preminger* (2015). Case in Sakha: Are two modalities really necessary? Natural Language and Linguistic Theory, Vol. 33(1), pp. 231–250.
32. *Lyutikova, Ekaterina* (2014). Case and noun phrase structure: differential object marking in Mishar dialect of Tatar [Padezh i struktura imennoy gruppy: variativnoye markirovaniye ob'ekta v misharskom dialekte tatarskogo yazyka], Rhema [Rhema. Rema], Vol. 4, pp. 50–70.
33. *Lyutikova, Ekaterina, and Asya Pereltsvaig* (2015). Noun phrase structure in articleless languages: universality and variation [Struktura imennoy gruppy v bezartiklevyx yazykax: universal'nost' i variativnost'], Topics in the study of language [Voprosy yazykoznaniya], Vol. 3, pp. 52–69.
34. *Lyutikova, Ekaterina, Sergei Tatevosov, Mikhail Ivanov et al.* (2006). Event structure and verb semantics in Karachay-Balkar [Struktura sobytiya i semantika glagola v karachaevo-balkarskom yazyke]. Moscow: IMLI RAN.
35. *Malchukov, Andrei, and Peter de Swart* (2008). Differential Case Marking and Actancy Variation. In: A. Malchukov and A. Spencer (eds), Handbook of Case. Oxford University Press, pp. 339–355.
36. *Marantz, Alec* (1991). Case and licensing. In Germán Westphal, Benjamin Ao, and Hee-Rahk Chae (eds.). Eastern States Conference on Linguistics, University of Maryland, Baltimore: Ohio State University, pp. 234–253.
37. *Mazuka, Reiko, and Barbara Lust* (1990). On parameter setting and parsing: Predictions for cross-linguistic differences in adult and child processing. In L. Frazier & J. De Villiers (eds.), Language Processing and Language Acquisition. Kluwer Academic Publishers, Dordrecht, pp. 163–206.
38. *McFadden, Thomas* (2004). The position of morphological case in the derivation: A study on the syntax-morphology interface. Doctoral dissertation, University of Pennsylvania.
39. *Mithun, Marianna* (1991). Active/agentive case marking and its motivations. Language, Vol. 67 (3), pp. 510–546.
40. *Moravcsik, Edith* (1974). Object-verb agreement. In Working papers on language universals, Vol. 15, pp. 25–140.

41. *Moravcsik, Edith* (1978). Agreement. In J. Greenberg (ed.). Universals of human language IV: syntax. Stanford, CA: Stanford University Press, pp. 331–374.
42. *Øvrelid, Lilja* (2008). Linguistic features in data-driven dependency parsing. CoNLL 2008—Proceedings of the Twelfth Conference on Computational Natural Language Learning, pp. 25–32.
43. *Øvrelid, Lilja, and Joakim Nivre* (2007). WhenWord Order and Part-of-Speech Tags are not Enough—Swedish Dependency Parsing with Rich Linguistic Features. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP), pp. 447–451.
44. *Payne, Doris, and Immanuel Barshi* (eds.) (1999). External possession. Amsterdam: John Benjamins.
45. *Pereltsvaig, Asya, and Ekaterina Lyutikova* (2014). Possessives within and beyond NP: Two ezafe-constructions in Tatar In: A. Bondaruk, G. Dalmi and A. Grosu (eds.). Advances in the Syntax of DPs: Structure, agreement, and case. Amsterdam: Benjamins, pp. 193–219.
46. *Preminger, Omer* (2011). Agreement as a fallible operation. Doctoral dissertation, Cambridge, MA: MIT.
47. *Preminger, Omer* (2014). Agreement and its failures. Linguistic Inquiry Monograph 68. Cambridge, MA: MIT Press.
48. *Santesteban, Mikel, Martin J. Pickering, Itziar Laka, and Holly P. Branigan* (2015). Effects of case-marking and head position on language production? Evidence from an ergative OV language. Language, Cognition and Neuroscience, Vol. 30, pp. 1175–1186.
49. *Seeker, Wolfgang, and Jonas Kuhn* (2013). Morphological and Syntactic Case in Statistical Dependency Parsing. Computational Linguistics, Vol. 39 (1), pp. 23–55.
50. *Shibatani, Masayoshi* (1994). An integrated approach to possessor raising, ethical datives and adversative passives. Proceedings of the Twentieth Annual Meeting of the Berkeley Linguistics Society. Berkeley, California, pp. 461–487.
51. *Silverstein, Michael* (1976). Hierarchies of features and ergativity. In Grammatical categories in Australian languages, ed. by R. M. W. Dixon. New Jersey, NJ: Humanities Press, pp. 112–171.
52. *Szabolcsi, Anna* (1983). The possessor that ran away from home. The Linguistic Review, Vol. 3, pp. 89–102.
53. *Testelets, Yakov* (2014). Structure of nominal constructons and the problem of case alternation in Adyghe [Struktura imennykh konstrukciy i problema cheredovaniya padezhey v adygskikh yazykakh]. In Daniel, Mikhail, Ekaterina Lyutikova, Vladimir Plungian et al (eds.). Language. Constants. Variables. In memoriam of Aleksandr Kibrik [Yazyk. Konstanty. Peremennyye. Pamyati Aleksandra Evgenyevicha Kibrika]. Sankt-Petersburg: Aleteya, pp. 536–551.
54. *Tsarfaty, Reut, Djamé Seddah, Sandra Kübler, and Joakim Nivre* (2013). Parsing morphologically rich languages: Introduction to the special issue. Computational Linguistics, Vol. 39 (1), pp. 15–22.
55. *von Heusinger, Klaus, and Jaklin Kornfilt* (2005). The case of the direct object in Turkish: Semantics, syntax and morphology. Turkic Languages, Vol. 9, pp. 3–44.

56. *von Heusinger, Klaus, Udo Klein, and Peter de Swart* (2008). Variation in differential object marking. Workshop on Case Variation, Stuttgart, June 2008. [URL: http://www.ilg.uni-stuttgart.de/projekte/C2/events/08CaseVariation/CaseVariationPaper.pdf]
57. *Woolford, Ellen* (2006). Lexical Case, Inherent Case, and Argument Structure. Linguistic Inquiry, Vol. 37, pp. 111–130.
58. *Yamashita, Hiroko* (1997). The Effects of Word-Order and Case Marking Information on the Processing of Japanese. Journal of psycholinguistic research, Vol. 26(2), pp. 163–188.

GRAMMATICAL DICTIONARY GENERATION USING MACHINE LEARNING METHODS

Mazurova M. (sleepofnodreaming12@gmail.com)

Ashmanov & Partners, Moscow, Russia

For the last decade, grammatical dictionaries have become not only a thing of theoretical value but an essential tool used in many fields of applied linguistics. However, the procedure of manual creation of a grammatical dictionary remains time- and labor-consuming. In this paper, the two-stage algorithm of automatic dictionary compilation, not requiring annotated texts, is proposed. As the source data, this system requires a formalized grammar description and a frequency distribution of a relatively large (hundred thousand tokens) corpus. Extending the principles commonly applicable to Indo-European languages, the research focuses on machine learning methods of corpora-based dictionary formation. Four machine learning models—SVM, random forest, linear regression and perceptron—are tested on the material of four languages: Albanian, Udmurt, Katharevousa, and Kazakh, and compared to a heuristic approach. While the linear models proved to be ineffective, other models' results were more promising: in an experiment with training and test sets formed from the same language's material, random forest reached 63% F-score, and SVM's results were also overdoing the baseline, however, the random forest model was unsuccessful. The best classifier in case of training and test sets based on the material of different languages was SVM. As a by-product of the experiments, the restrictions on the input were postulated: the approach 'as is' is not applicable to languages where inflections are strongly homonymic, and, on the contrary, is promising applied to an agglutinative language.

Keywords: grammatical dictionary, morphology, machine learning, morphological analyzer

ГЕНЕРАЦИЯ ГРАММАТИЧЕСКОГО СЛОВАРЯ ДЛЯ ПРОИЗВОЛЬНОГО ЯЗЫКА С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Мазурова М. (sleepofnodreaming12@gmail.com)

«Ашманов и партнеры», Москва, Россия

Ключевые слова: грамматический словарь, морфология, морфологический анализ, машинное обучение

1. Introduction

Grammatical dictionary is a formalized description of lexemes' word formation in a language. Over the last decade grammatical dictionaries have turned out to be relevant to many tasks of modern applied linguistics. In NLP they are an essential part-of-speech (POS) taggers, spellcheckers, etc. For example, POS-tagger MyStem [12] uses a dictionary arranged as a trie of suffixes and a set of stem tries for checking if there is a canonical tagging option for a token being processed.

For Russian, there is an outstanding grammatical dictionary: 'Russian Grammar Dictionary' by A. Zaliznyak, an exhaustive inventory of Russian affixation put together in the late 1970s. It has inspired many Slavic researchers to create similar dictionaries, for example, for Polish [11] and Bulgarian [6]. As for non-Indo-European languages, there exists a Bashkir dictionary [1], but this work demonstrates a series of faults, making it difficult to use and potentially inappropriate for NLP purposes [9]. All the dictionaries listed above were compiled manually, and compilation of a grammatical dictionary is a complex time-consuming task that requires expertise of experienced linguists. In order to make the task of compiling grammatical dictionaries less problematic, an automatic approach can be suggested. In this paper I aim to develop an approach to automatic dictionary generation which does not require annotated texts and is suitable for relatively underresourced languages.

2. Related Studies

Dictionary extension methods for Russian were discussed in [13] by Segalovich & Maslov. The authors postulate a principle: 'the paradigm must be built based on corpus data'; according to it, lexeme's paradigm is a set of all its word forms found in a corpus. A program should find lexeme hypotheses using a list of suffixes and following the 'corpus-based paradigm' principle, and then filter the data with a series of heuristics. Performance analysis of the algorithm was conducted later by O. Lyashevskaya [7]. The author, however, applies another set of filters; 'the longest stem heuristics' is proposed. The similar heuristics is used in [5] for Czech data (it is worth noting that earlier Hana & Feldman [4] used the opposite approach for a similar purpose). Later, a completely different approach to the problem is applied: for example, 'Bystroslovar' is generated from a large data array with the use of machine learning methods [14].

3. Dictionary Draft Generation

The first stage was developing a draft generation algorithm. Following the idea of the Porter's stemmer [10], the system uses a list of inflections to divide a word form into an inflection and a stem, and then combines an appropriate set of inflection-stem pairs into a lexeme as proposed in [13]. Note that a notion 'stem' (and, consequently, 'inflection') here is not linguistically correct: a stem is defined as an unchangeable part of a set of word forms of a lexeme (in the most common case, a common part of all its word forms), while an inflection is a changeable one.

3.1. Algorithm & Implementation

The system works with frequency distribution of word forms of a corpus which contains more than several hundred thousand words. The system also requires a formal description of a language's morphology in UniParser format [2].

As was mentioned earlier, the main procedure is stemming-like: the system parses a word form into a stem and an inflection, providing all parsing options allowed by the grammar used. Let us consider an example from Greek: there is an inflection '*.α*' in a verbal paradigm, and an inflection '*.ότερα*' in an adjective paradigm. Following the above rule, the system, given a word form '*ειδικότερα*', generates two parsing options:

(1) *ειδικότερ. + .α*
ειδικ. + .ότερα

Having produced the set of parsing hypotheses, the system attributes each of them to possible paradigm types checking which paradigm has the given affix; all the hypotheses are saved to a data storage, accompanied with a word form frequency. If there are stem alternations in the grammar, then lexeme joining is executed: the system generates alternatives for each stem found and, if there is a stem generated in the storage, puts the lexeme parts together.

The algorithm survived several different implementations [8], and the latest and the fastest one[1] is based on a custom finite-state automaton. Reading the grammar, the system forms a list including all paradigms' affixes, and then compiles an NFA capable to find multiple substrings. The NFA is used to quickly get all the parsing hypotheses, which are attributed to paradigms later by means of a structure mapping every inflection's graphical representation to a set of its attribution options.

3.2. Performance

Below I provide a theoretical estimation of processing time. The time of NFA compilation is limited by the time necessary to read all inflections of a grammar; the compilation is executed once and its time is negligible compared to the time required for the preceding UniParser grammar compilation which is rather time-consuming.

The time required to process one word form is bounded above by an exponential function of its length because the automaton is non-deterministic; however, a possible length of a word form is limited by a small number, so the complexity may be considered asymptotically constant. As a result, the only parameter affecting the processing time is the length of an input, e.g. it is O(N), where N is a number of graphically unique word forms in the input.

[1] https://github.com/sleepofnodreaming/gramdicmaker2016

Below (Table 1) I provide empirical performance measurements[2] on Kazakh data. Generally, the Kazakh corpus, 904,561 tokens in size, contains 107,704 unique word forms; for this study the word list was divided into four portions to check how the processing time increases if the input gets bigger. The number of paradigms is the same for the series of tests and equals 28.

Table 1. Performance (depending on the input frequency distribution size)

FD size, word forms	Time, s
26,926	1.855
53,852	3.523
80,778	6.797
107,704	7.005

The next question was how the processing time depends on a number of inflections in a NFA. This time, the number of words was fixed: the full frequency list was used. However, the number of affixes in an automaton varied. As adding separate inflection to the automaton is not allowed, the number of inflections in process was changed by varying the number of paradigms.

Table 2. Performance (depending on the number of inflections)

Paradigms	Number of affixes	Time, s
N-soft	766	2.01
All nominal	3,104	4.29
All verbal	82,260	5.85

Multiplication of a size of an input grammar does not lead to rapid growth of the processing time (see Table 2). This behavior of the system does not contradict the theoretical presuppositions made above.

4. Data Filtering

If the grammar used for the system is correct, a draft generated with the use of the described procedure includes all lexemes present in the source corpus. However, a considerable percentage of the formed lexemes are products of misparsing. For instance, example (2) was attributed to a verbal paradigm, but this set of word forms corresponds to a Kazakh noun *сұлуысың* '*beauty*':

[2] The machine all measurements are made on has 8Gb RAM, Intel Core i5 2,7GHz CPU, and runs MacOS 10.10.

(2) *сұлуы.*

.	*imper,2,sg / indic,prs,3*
.мыз	*indic,prs,1,pl*
.сың	*indic,prs,2,sg*

Let us call all the dictionary draft units—both real lexemes and mistakenly formed ones—pseudolexemes. In [8], a simple frequency-based heuristic classifier was used to remove false lexeme units. The threshold was set as an empirically chosen logarithmic function of a number of word forms of a lexeme: if general pseudolexeme frequency is more than a threshold value, it is considered a real one. There were also two additional thresholds: if a number of word forms of a pseudolexeme is less than the threshold-min, it is always removed; if a number of word forms is more than the threshold-max, a pseudolexeme is never removed.

Let us consider the above classifier baseline; however, in this paper I study the other approach—machine learning binary classification based on lexeme's distributional features. I also research an opportunity to use a dataset formed from another language's data.

4.1. Data Sets

For the current study I used data from the project 'Corpus Linguistics' (Udmurt, Albanian, Katharevousa) and Almaty Corpus of Kazakh[3]; besides the corpora, the source data included UniParser grammars and dictionaries[4] providing the information about lexeme's POS and paradigm type (see Table 3).

Table 3. Corpora & Dictionaries

Language	Corpus size, number of word usages	Dictionary size, lex
Katharevousa	359,805	403 (adjectives only)
Udmurt	6,368,427	21,656
Albanian	19,543,008	45,861
Kazakh	904,561	22,024/14,527

Using these sources four dictionary drafts were compiled: for Kazakh, the draft included verbal and noun pseudolexemes; for Udmurt, there were noun, verbal and adjective lexemes, and Albanian and Katharevousa draft dictionaries consisted of adjectives. Then the drafts were annotated automatically with the use of the dictionary data; no manual revision was made. Annotation was made according to the following principle: a pseudolexeme is a real lexeme if, first, all its stems postulated are a part of a lexeme found in a dictionary and, second, there is no exceeding stem in a lexeme. As a result, four data sets were formed.

---

[3] http://web-corpora.net/

[4] The Katharevousa dictionary was formed automatically with the use of the frequency filter and manually reviewed. For this reason, the dictionary lacks low-frequency lexemes.

Table 4. Data sets

Language	Lexemes formed		Valid lexemes		Invalid lexemes		Valid lexemes, %	
	Full	Cut	Full	Cut	Full	Cut	Full	Cut
Albanian	2,047,093	799,004	5,635	4,378	2,041,458	794,626	0.27	0.54
Kazakh	62,704	23,354	7,479	5,155	55,225	18,199	11.9	22.1
Katharevousa	3,959	—	370	—	3,589	—	9.3	—
Udmurt	278,036	101,577	7,728	5,789	270,308	95,788	2.7	5.7

Generally, a proportion of valid lexemes is not big: from language to language, it varies from 0,27% to 11,9%; but, in general, it exceeds error rate (see Table 4). However, in the case of Albanian it is abnormally low because of the language's extensive homonymy rate.

As the percent of valid lexemes turned out to be quite low and the threshold-min filter proved to be effective in [8], I formed three additional data sets, removing lexemes with frequency under 5. The thresholded data sets contain more valid lexemes, but in the case of Albanian thresholding did not solve the problem: the percentage of valid Albanian lexemes still turned out to be extremely low.

4.2. Features

The next step would be choosing a set of distributional features. Let c be a grammatical category that is represented with inflections of a paradigm π. Frequencies of c's values define probability distribution inside π: I will call it $d(c,\pi)$. Respectively, distribution of c's values inside a pseudolexeme w is called $d_w(c,w)$. Although it is probable that distributions $d(c,\pi)$ vary dramatically from paradigm to paradigm and from language to language, it is reasonable to assume that some statistical features of a real lexeme's $d_w(c,w)$ are similar to features of $d(c,\pi)$. The set of features following the hypothesis is:

1. average entropy of $d_w(c,w)$ for the paradigm's c's, c size normalized;
2. minimum entropy of $d_w(c,w)$ for the paradigm's c's, c size normalized;
3. variance of a distribution of all entropies of $d_w(c,w)$.

Being guided by a research conducted by A. Sokirko on Russian material [14], I added a series of features based on the completeness rate of of a formed lexeme:

4. percentage of lexeme's word forms found in a corpus (graphically identical word forms are considered one form);
5. percentage of lexeme's grammatical forms found.

The next presumption is the following: if a pseudolexeme results from misparsing due to cross-paradigm homonymy, distribution of category values inside a pseudolexeme is often affected. In Katharevousa, for instance, adjective suffixes are homonymic

to noun suffixes, so a noun may be interpreted as an adjective. However, an adjective pseudolexeme formed this way is going to lack the majority of word forms: a noun supposedly covers word forms of the same gender. As it is proposed in [13], these cases may be handled heuristically, but in my study I transform this idea into the following feature:

6. number of pseudolexeme's categories having the only value.

Other features that are not based on distribution of categories are:

7. entropy of word forms' frequencies, divided by a size of a paradigm;
8. word forms' frequency distribution entropy, not normalized;
9. different word form number—lexeme occurrence ratio.

5. Evaluation Methods

For evaluating the quality of a dictionary cleaning standard relevance measures were used: precision (P), recall (R) and F score. Additionally, frequency-weighted measures were set up:

$$R_f = \frac{\sum_{f \in \pi} \sum_{\omega \in tp} \sum_{s \in \omega} I(f,s)}{\sum_{f \in \pi} \sum_{\omega \in p} \sum_{s \in \omega} I(f,s)}$$

$$P_f = \frac{\sum_{f \in \pi} \sum_{\omega \in tp} \sum_{s \in \omega} I(f,s)}{\sum_{f \in \pi} \sum_{\omega \in t} \sum_{s \in \omega} I(f,s)}$$

$$F_f = \frac{2 \cdot P_f \cdot R_f}{P_f + R_f}$$

The abbreviations in the formulas above are: tp—true positive results, tn—true negative results; p и n are numbers of positive and negative results, respectively; $I(f,s)$ is a function defining a number of tokens that may be represented as a combination of an inflection f and a stem s in a corpus.

5.1. Extension of an Existing Dictionary

In the beginning of a series of experiments, an analysis of relevance of the different features was conducted, in order to form a set of features that would be suitable for extension of an existing dictionary. Four supervised ML models were tested: SVM, linear regression, perceptron and random forest. Four training sets were used: the Kazakh and the Albanian ones, both full and thresholded. To evaluate the results, cross-validation was conducted: two-fold for Albanian case and four-fold for Kazakh case.

Table 5. Classification of Kazakh pseudolexemes

Model	Dataset	P	P_f	R	R_f	F	F_f
Perceptron	full	0.899	0.732	0.011	0.072	0.022	0.131
Perceptron	cut	0.729	0.729	0.045	0.094	0.085	0.167
Linear Regression	full	0.956	0.617	0.001	0.026	0.200	0.050
Linear Regression	cut	0	0	0	0	0	0
SVM	full	0.340	0.592	0.03	0.259	0.055	0.360
SVM	cut	0.316	0.591	0.03	0.259	0.055	0.360
Random Forest	*full*	0.540	*0.669*	*0.299*	*0.525*	*0.385*	*0.589*
Random Forest	cut	0.505	0.701	0.310	0.571	0.384	0.629
Base line	—	*0.333*	*0.410*	*0.591*	*0.939*	*0.426*	*0.571*

In this experiment (see Table 5), linear classifiers (LR and perceptron) proved to be ineffective. The performance of other classifiers is much better, but they did not outdo the heuristic one: the F score of the best ML classifier is 38.5%, while the baseline result is 42.6%. However, the weighted results are different: random forest showed to be better than all other classifiers, with F score equal to 62.9%. As for SVM, it demonstrated satisfactory weighted results, although the unweighted were poor: the weighted F score was about 36% vs. 5.5% in the unweighted case.

I will further discuss the problem with the Albanian set (see Table 6). All the classification methods used were extremely ineffective: none of the ML classifiers reached at least 6% precision, while the maximum recall was 21%. This effect can be explained by the data characteristics: the share of real lexemes is lower than the noise rate can normally be, and, as a result, the data are supposed to be not filterable with the use of any ML. To estimate the scale of the problem, I address a specific example: a noun *ngarkesë 'load'* was attributed to each of 16 adjective paradigms specified in the grammar; although these pseudolexemes consist of one word only, a number of different grammatical forms listed is sufficient:

Table 6. Classification of Albanian pseudolexemes (cut data set)

Model	P	R	F
Perceptron	0.039	0.21	0.0658
Linear Regression	0	0	0
SVM	0	0	0
Random Forest	0.053	0.003	0.0057

(3) *ngarkes.*

.e	*f.sg / f.pl / f.sg.nom.indef / f.sg.acc.indef / f.pl.nom.indef / f.pl.acc.indef*

Much more successful processing of the Kazakh data is caused by agglutinative grammatical system features: in this case, an inflection, being a combination of affixes, tends to be less homonymic. As a result, a set of grammatical categories present in a misparsed lexeme often differs dramatically from the correct one:

(4) *бағана.*

.	*imper,2,sg / indic,prs,3*
.сын	*imper,3*
.ғы	*opt1,3*

5.2. Classification of Another Language's Data

At the next stage of the research, I tested whether it is possible to train a classifier on a data set formed from another language's material. This approach is not widespread but is used for some purposes: in [4], another language's data are used to train a HMM, and a possibility to train a morphological analyzer is postulated in [3]. As test sets, data from languages of two different types of inflection, Udmurt (agglutinative) and Katharevousa (cumulative) were used. I aimed to test the following hypothesis: it is more effective to use a training set formed of morphologically similar language's material.

Unfortunately, Kazakh turned out to be the only training set suiting for this experiment. A set of ML models was the same as in the previous experiment, but perception and linear regression's results remained extremely unsuccessful, and I will not discuss them further.

In this experiment, the use of random forest proved to be also inappropriate: in Katharevousa case, the model does not work, considering almost all the lexemes misformed; as for Udmurt, the results are also poor: the precision is under 10%, and the recall is under 15%. However, this result is natural: the forest adapts the training data strongly, and the classification is based on certain values of a feature.

Taking a closer look on the classification results, I can see that a typical Katharevousa lexeme approved by the ML consists of a relatively big number of word forms (and, consequently, looks more like a Kazakh one). For instance, a Katharevousa lexeme *πληγή* '*wound*' (nine different forms are found) was considered a well-formed adjective, although it obviously lacks neutral and masculine forms:

(5) *πληγ.*

.άς	*pos,f,pl,acc*
.ή/.ὴ	*pos,f,sg,nom*
…	
.ῶν	*pos,pl,gen*
.ὰς	*pos,f,pl,acc*

Additionally, I studied weights of features for the classifier trained on the Kazakh data (see Fig. 1). The most significant features are #6 (a number of categories having the only value) and #8 (entropy, not normalized). The contribution of the majority of features is equally moderate (about 10%), and the only feature (minimum entropy of (*c*,*w*) for the paradigm's categories) proved to be useless, contributing almost nothing.

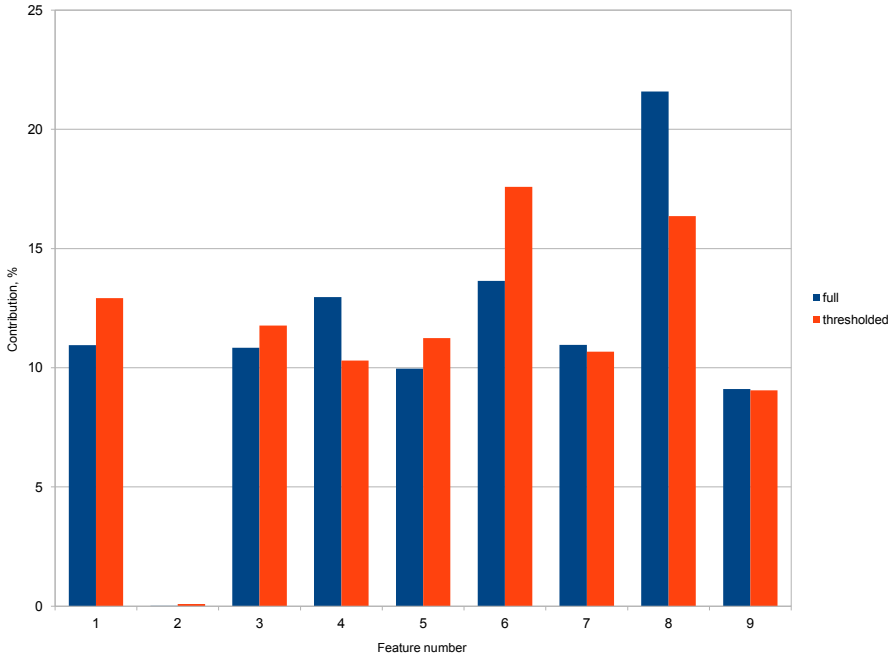


Fig. 1. Feature weights, random forest

As for SVM results, they are moderate but acceptable and still in accordance with the initial hypothesis: the precision is higher than the baseline for Udmurt, which is more similar to Kazakh morphologically. The weighted recall for Udmurt reaches 74%, and the precision is about 30%; it results in 42% F score. In the Katharevousa case the results are poorer: the maximum weighted precision and recall are 7.3% and 47.6%, respectively.

6. Conclusions

In this paper, I proposed the two-stage algorithm of grammatical dictionary generation / extension for any language. The first stage implementation, draft generation, turned out to be effective enough. As for the second stage, filtering, four ML models were tested, and two of them performed successfully. Generally, the results are moderate but promising: although the approach proposed does not work with languages

featuring rich cross-paradigm homonymy, it proved to be perspective, outdoing the baseline filter both in existing dictionary extension and, conditionally, brand new dictionary generation case: it is likely that the use of the another language's training data is admissible if its inflectional model is similar to a model of a language to be processed. On the other hand, the insufficient amount usable data prevents me from conducting more detailed experiments: the dictionaries used are not completely correct from the point of view of computational linguistics, and it presumably affects the quality of results, and the number of available data sets is limited.

Acknowledgement

I am grateful to Danya Alexeyevsky, Boris Orekhov, and Andrey Kutuzov for their consultations at different stages of my research.

References

1. *Akhtyamov, M.* (1994), Bashkir Grammar Dictionary. Inflection [Башҡорт теленең грамматика һүҙлеге. Һүҙүҙгәреше]. 'Bashkortostan', Ufa.
2. *Arkhangelskiy, T. A.* (2012), Principles of development of a morphological analyzer for languages with different structures [Printsypy postroyeniya morfologicheskovo parser dlya raznostrukhturnykh yazykov], Moscow.
3. *Brants, T.* (2000), A statistical Part-of-Speech tagger, Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLP-2000), Seattle, WA, USA, pp. 224–231.
4. *Hana, J., & Feldman, A.* (2004), Portable language technology: Russian via Czech, Proceedings of the Midwest Computational Linguistics Colloquium, Bloomington, Indiana.
5. *Kanis, J., & Müller, L.* (2005), Automatic lemmatizer construction with focus on OOV words lemmatization, Text, speech and dialogue, pp. 132–139.
6. *Koeva, S.* (1998), Bulgarian Grammatical dictionary. Organization of the language data, Bulgarian language, Vol. 6, pp. 49–58.
7. *Lyashevskaya O. N.* (2007), Towards the lemmatization of word forms absent from dictionary [K probleme lemmatizatsii neslovarnykh slovoform], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2007" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2007"], Bekasovo, pp. 407–412.
8. *Mazurova, M.* (2014), Generating a farmalized description of a language's lexis from unannotated texts [Porozhdeniye formalizovannogo opisaniya leksiki yazyka na osnove tekstov], Moscow.
9. *Orekhov, B. V.* (2014), Problems of grammatical annotation of texts in Bashkir [Problemy morfologicheskoy ravmetki bashkirskih tekstov], Proceedings of TEL-2014 [Trudy Kazanskoy shkoly po kompjuternoy i kognitivnoy lingvistike TEL-2014], Kazan, pp. 135–140.

10. *Porter, M. F.* (1980), An algorithm for suffix stripping, Program, Vol. 14(3), pp. 130–137.
11. *Saloni, Z., Gruszczyński, W., Woliński, M., & Wołosz, R.* (2007), Grammatical Dictionary of Polish, Studies in Polish Linguistics, Vol. 4, pp. 5–25.
12. *Segalovich, I.* (2003), A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine, MLMTA, Los Angeles, pp. 273–280.
13. *Segalovich, I., Maslov, M.* (1998), Russian morphological analysis and synthesis, generating inflectional models for word absent from dictionary [Russkiy morfologicheskiy analiz i sintez s generatsiyey modeley slovoizmeneniya dlya ne opisannykh v slovare slov]. Dialogue'98 [Dialog 98], Kazan, Vol. 2, pp. 547–552.
14. *Sokirko A. V.* (2010), Bystroslovar': morphological prediction of new Russian words using very large corpora [Bystroslovar': predskazanie morfologii russkikh slov s ispol'zovaniem bolshikh lingvisticheskikh resursov], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2010" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2010"], Bekasovo, pp. 450–456.
15. *Zaliznyak A.* (1977), Russian Grammar Dictionary [Grammaticheskij slovar' russkogo jazyka], Moskva.

POSSESSIVES IN PARALLEL ENGLISH-CZECH-RUSSIAN TEXTS

Nedoluzhko A. (nedoluzko@ufal.mff.cuni.cz)
Charles University in Prague, Prague, Czech Republic

Schwarz (Khoroshkina) A. (annakhor@gmail.com)
ABBYY; Moscow State University, Moscow, Russia

Novák M. (mnovak@ufal.mff.cuni.cz)
Charles University in Prague, Prague, Czech Republic

We present a corpus-based analysis of the use of possessive and reflexive possessive pronouns in a newly created English-Czech-Russian parallel corpus (PCEDT-R). Automatic word-alignment was applied to the texts, which were subsequently manually corrected. In the word-aligned data, we have manually annotated all correspondences of possessive and possessive reflexive pronouns from the perspective of each analysed language. The collected statistics and the analysis of the annotated data allowed us to formulate assumptions about language differences. Our data confirm the relative frequency of possessive pronouns in English as compared to Czech and Russian, and we explain it by the category of definiteness in English. To confirm some of our hypotheses, we used other corpora and questionnaires. We compared the translated texts in Czech and Russian from our corpus to the original texts from other corpora, in order to find out to what degree the translation factor might influence the frequency of possessives.

Key words: possessive pronouns, possessive reflexive pronoun, coreference, comparative analysis, parallel corpus, English, Czech, Russian

ПРИТЯЖАТЕЛЬНЫЕ МЕСТОИМЕНИЯ В ПАРАЛЛЕЛЬНЫХ АНГЛО-ЧЕШСКО-РУССКИХ ТЕКСТАХ

Недолужко А. (nedoluzko@ufal.mff.cuni.cz)
Карлов университет в Праге, Прага, Чехия

Шварц (Хорошкина) А. (annakhor@gmail.com)
ABBYY; МГУ, Москва, Россия

Новак М. (mnovak@ufal.mff.cuni.cz)
Карлов университет в Праге, Прага, Чехия

Ключевые слова: притяжательные местоимения, возвратно-притяжательное местоимение, кореферентность, сравнительный анализ, параллельный корпус, английский язык, чешский язык, русский язык

1. Introduction

Possessivity and how it is expressed by means of possessive and reflexive possessive pronouns is a popular topic in theoretical linguistics. This subject is discussed from many different perspectives, e.g. typological, historical, semantic, syntactic and discursive. The approach used in this paper is contrastive, multilingual and corpus-based.

While analysing the correspondences between coreferential expressions in parallel English-Czech texts (Novák—Nedoluzhko, 2015), we have observed that possessive pronouns in English appear more frequently than they do in the same positions in Czech. This led us to carry out a systematic analysis of such cases. While analysing the Czech counterparts we discovered many language-specific details in Czech that required a comparison to another Slavic language in order to properly interpret this more thoroughly. And we chose Russian[1] for this analysis. The use of possessives has been analysed in detail for each language separately (see Section 2). However, Czech and Russian are typologically and genetically very close, and the rules for the use of possessives are quite similar. So, the general rule both for Czech and Russian is that a reflexive possessive can substitute personal possessive pronouns if it is coreferential with the subject. On the other hand, the use of possessives is not completely identical. For example, the distributive meaning of the reflexive possessive *svoj* is very common in Russian, while in Czech it is completely marginal, see Example 1:

(1) RU: *U každogo učenogho jesť svoja biblioteka*—CZ: *Každý vědec má *svou/vlastní knihovnu* [EN: Each scientist has his own library.]

The main concern of this paper is to investigate the use of possessive and reflexive possessive pronouns in English, Czech and Russian. For our analysis, we used a newly created three-language Czech-English-Russian parallel corpus (PCEDT-R, see Section 3), which we provided with automatic word alignment, its manual correction and annotation. For the aligned annotated data, we compiled the statistics of correspondences of the pronouns under analysis in the three languages (Section 4). The next step was to analyse the resulting correspondences. In Section 5, we provide proof for the assumption that Russian possessive and reflexive possessive pronouns occur less frequently than in English, however we also show that they are used significantly more frequently than in Czech. The differences between Czech and Russian are addressed in Section 6.

Overall in this paper, we use existing theoretical knowledge from non-corpus-based monolingual research, the annotation of corpus data and our language intuition to formulate hypotheses about the use of possessives in Czech, English and Russian and support them with the statistics from the three-language parallel corpus analysis. We have not found the theoretical basis for all of our assumptions yet, which is planned for the next stages of the research. We believe that the research in this field is helpful both for the improvement of machine translation work (e.g. it would be easier to identify which English possessives should be translated into Czech and Russian and which should not) and for theoretical comparative language analysis.

[1] One of the reasons is that there is extensive literature on this topic in Russian.

2. Related work

There is a variety of means to express the notion of possessivity (see e.g. a detailed survey in Brykina 2009). In this paper, we are interested, above all, in pronouns. In English, there is one group of possessive pronouns (*his, her, its, their*)[2], which are translated into Czech and Russian as possessive (*jeho, její, jejich* in Czech, *jego, jeje, ich* in Russian) and possessive reflexive (*svůj* in Czech and *svoj* in Russian) pronouns.

For Czech, the description of personal and possessive reflexive pronouns begins with Svoboda (1880) and is further addressed in a number of theoretical studies and grammars (see e.g. in Gebauer 1890, Trávníček 1951, Daneš—Hausenblas 1962, Panevová 1986, Dočekal 2000, etc.). The concurrence of possessive and reflexive possessives is described in most detail for syntactic constructions with one predication (*Já čtu svou/?mou knihu = I'm reading my book*) and for cases with an explicitly embedded predication (*Já slyším tě zpívati svou/mou/tvou oblíbenou píseň. = I hear you sing your/my favorite song*[3]). Many examples of different types are systematically analysed, but the results are formulated rather as tendencies than as strong rules and are based on syntactic (Trávníček 1951, Daneš—Hausenblas 1962, Dočekal 2000), semantic (Panevová 1986) or discursive (Čmejrková 1998) criteria. For Russian, the concurrence of possessive and reflexive possessive pronouns is addressed in most detail in Padučeva (1985); the author provides ten distinctions between the different types and explains them with syntactic, semantic and referential arguments. Some non-typical types of control of possessive reflexive were addressed in Fed'ko (2007).

As for the meanings of the reflexive possessive in Russian, Padučeva (1983) considers 6 different types of *svoj*. The basic type is the possessive form of a reflexive pronoun *sebja* (*oneself*), the same as in Czech; other meanings are derived from the basic one with additional functions added, such as the distributive meaning, the meaning of 'special', 'appropriate', the contrast 'one's own' ↔ 'somebody's else' and so on. In contrast to Czech (Panevová, 1986), Padučeva describes the tendency of *svoj* to take part in different kinds of idiomatic expressions. And in addition to the six meanings of *svoj* introduced in Padučeva (1983), Brykina (2009) mentions that it may be used in sentences where it is semantically redundant, though it also possesses discursive or syntactic functions, such as indicating the focus of attention and maintaining referential connectivity (Brykina, 2009:135ff.).

The study of meaning and distribution of possessive and reflexive possessive pronouns in our analysis is closely connected to the subject of external possessivity. In this respect, we took into account the research provided in Brykina (2005, 2009) and Kibrik (2003) for Russian and Krivan (2007) for Czech.

---

[2] Here and later, we only speak about possessives in the third person.

[3] This Gebauer's (1890) example was used by many researchers addressing this topic.

As for existing parallel corpora including all the languages under analysis, there are the Intercorp (a part of the Czech National Corpus)[4] and the ParaSol[5] multi-language corpora. In Intercorp, parallel data can be excerpted for pairs of languages, namely separately for e.g. English-Czech and Czech-Russian. Texts in both corpora are automatically sentence-aligned, there is no word-alignment. As far as we know, parallel language data have not been used for this kind of analysis yet. The research addressing semantic and pragmatic functions of possessives in Czech mostly relies on the linguistic intuition of the authors and the analysis of constructed or specially found examples.

3. Data and methods

Our core analysis is based on the newly created three-language parallel English-Czech-Russian corpus. The English-Czech part of it was taken from the Prague Czech-English Dependency Treebank (PCEDT, Hajic et al. 2012) and translated into Russian (in what follows, the abbreviation **PCEDT-R** will be used for this three-language parallel corpus). Given the size of PCEDT, the translation and the manual annotation of word alignment of the entire PCEDT would be extremely time-demanding. We therefore limited the dataset to only 1078 sentences located in the first half of the PCEDT section 19, i.e. the 50 documents from wsj_1900 to wsj_1949.

The English part consists of the Wall Street Journal (WSJ) section of the Penn Treebank (Marcus et al., 1999). The Czech and Russian parts were manually translated from the English source sentence by sentence.[6] The linguistic annotation in the English-Czech part of PCEDT-R is provided within the following annotation layers: the lowermost "word" layer (w-layer) representing the tokenized plain text, the morphological layer (m-layer) containing automatic part-of-speech tagging and lemmatization, the analytical layer (a-layer) representing surface dependency syntax, and the deep syntax or tectogrammatical layer (t-layer). The t-layer includes semantic labeling of content words (nouns, adjectives, adverbs, and verbs) and coordinating conjunctions, ellipsis reconstruction, coreference annotation, and argument structure description based on a valency lexicon. PCEDT-R is the same excerpt of texts that was used for analysis of coreferential expressions in English and Czech (Novák—Nedoluzhko, 2015), thus it already contains manual annotation of word alignment for English per-

[4] Texts in Intercorp are taken from the Project Syndicate website (http://www.project-syndicate.org).

[5] The project of the Humboldt University of Berlin, http://www.slavist.de/

[6] The Czech translation has been created for the whole PCEDT (ca. 50,000 sentences, see Hajič et al., 2012) by professional translators. The translation into Russian has been completed recently by A. Schwarz. The Czech and Russian translators were instructed to keep the sentence structure of the source texts. The aim of the research was completely unknown to Czech translators. As for the Russian translator, she was informed that the texts were intended for a comparative study of coreference chains. This fact could affect the translation, therefore we compared the differences between translated texts in Russian and Czech. The results showing the smaller difference between original and translated texts for Russian than for Czech in our corpus are presented in Section 7 below.

sonal and possessive pronouns and Czech relative pronouns. The Russian part was automatically aligned with the Czech part of PCEDT using GIZA++ (Och and Ney, 2000), which was run on a large amount of parallel Czech-Russian data. The resulting triples containing possessive units (in at least one of the languages used) have been manually annotated and analysed from the perspective of each language separately.

Table 1 shows some of the basic statistics and information related to the present work calculated on PCEDT-R.

Table 1. Statistics for PCEDT-R

	English	**Czech**	**Russian**
texts			50
sentences			1,078
tokens	26,560	25,477	25,396
possessive	238	106	120
possessive-reflexive pronouns	—	91	85
morphological and syntactic annotation	yes	yes	no (planned)

To confirm our hypotheses formulated on the basis of examples and statistics from PCEDT and its Russian translation, we also used the examples from ***InterCorp***. However, parallel texts there only have sentence automatic alignment and automatic morphological tagging, they are not word-aligned. It means that searching for possessive pronouns there leads to numerous false results. In this study, we do not use the statistics obtained from this corpus. We used it to search for specific triples of sentences (ENG-CS-RU) for cases that we considered to be of interest. In general, over 600 English-Czech-Russian triples have been concerned in Intercorp, and only 70 of them are relevant for our analysis.

Our core analysis is based on corpus data, but to confirm some of our general estimates, we also used two short questionnaires for native speakers of Czech. Mainly, they contain questions aiming to find systematic differences between Czech and Russian possessives and are based on the functional description of reflexive possessive pronoun *svoj* in Russian completed by E.V. Padučeva (Padučeva, 1983).

4. Statistics

For PCEDT-R, we have calculated the number of occurrences of counterparts of aligned possessive and reflexive possessive pronouns. The calculation has been completed for each of the analysed languages.

Tables 2, 3 and 4 show the statistics for the aligned counterparts for English, Czech and Russian respectively. Possessives in language A may be mapped on the following categories in language B:

- possessive pronouns (the *poss* label, e.g. EN: *his*—RU: *jego*—CZ: *jeho*);

- reflexive-possessive pronouns (the *refl-poss* label, e.g. EN: *his*—RU: *svoj*—CZ: *svůj*);
- nouns, anaphoric zeros, demonstrative and personal pronouns (the *NP* label, see Example 2);

(2) EN: <*His*> [Steppenwoolf's—AN et al.] *board members alone have pledged $800,000.*—RU: *Только члены правления <Степпенвулф> обещали $800 000.*

- external possessive expressions, the definite article (in English) or relative clauses (the *other* label, see Example 3, where the possessive meaning is expressed in Czech with external dative reflexive *si*, and Example 8 in Section 5 below for the definite article);

(3) EN: *Glenn Beebe had sued the company after installing Burlington carpets in <his> office.*—CZ: *Glenn Beebe zažaloval společnost poté, co <si> koberce Burlington položil do kanceláře.*

Possessive and reflexive possessive pronouns may remain unaligned in two cases: Either when no possessive expression has been used in the same syntactic construction (the *no-poss* label), or the syntactic construction of the translated sentence has been reformulated, making the word alignment impossible (the *reword* label). Rewordings include cases where e.g. a biclausal construction in one language becomes a monoclausal construction in the other (see Example 9 in Section 5 below), comitative in one language and coordination in the other (*a boy with his father* vs. *a boy and his father*) and so on.

Table 2 shows that a significant[7] part of English possessive pronouns have other (or no) means of expression in Czech and Russian: Only 72% (92+80) and 82% (112+83) of English possessive pronouns are expressed by Czech and Russian possessive and reflexive-possessive pronouns, respectively. However, there is also a significant difference in expressive means between Czech and Russian. The reasons for such a difference (about 10%) can be found in the translator's style or special language features and will be addressed in more detail in later sections. It is also possible that the difference between the frequency of pronouns Russian and Czech is occasional, the lack of pronouns in Czech being partly compensated by rewording (15 in Czech vs. 10 in Russian) and externally expressed possessivity.

Table 2. Counterparts of English possessive pronouns in Czech and Russian

			external means		not aligned	
	poss	**refl-poss**	**NP**	**other**	**no-poss**	**reword**
CZ	92	80	6	6	39	15
RU	112	83	5	3	25	10

[7] This is significant at p-level $p \leq 0.05$. Significance has been calculated by bootstrap resampling using 100,000 samples. The same holds for all other claims about differences referred to as significant.

Table 3. Counterparts of Czech possessive and reflexive possessive pronouns in English and Russian

			external means		not aligned	
	poss	**refl-poss**	**NP**	**other**	**no-poss**	**reword**
EN	174	-	5	6	3	9
RU	94	62	9	—	22	10

Table 4. Counterparts of Russian possessive and reflexive possessive pronouns in English and Czech

			external means		not aligned	
	poss	**refl-poss**	**NPs**	**other**	**no-poss**	**reword**
EN	196	—	2	2	3	2
CZ	81	73	6	3	28	14

5. Analysis: English ↔ Czech & Russian

One of the most interesting points addressed in Novák—Nedoluzhko (2015) concerns the expression of possessivity in English and Czech. The statistics of the correspondence of English possessive pronouns to their Czech counterparts confirms the general tendency of Czech to express personal possessive pronouns less frequently than in English. For example, in Czech, it is not common to use a possessive (or a reflexive possessive) pronoun in sentences like (4). However, it is not ungrammatical. The Czech sentence in Example 4 would remain grammatically correct after adding a reflexive possessive *svůj*.

(4) EN: *As a result of <their> illness, they lost $1.8 million in wages and earnings.—*
CZ: *Důsledkem* (<své>) *nemoci přišli na mzdách a výdělcích o 1.8 milionu dolarů.*

We suggest that the high frequency of possessives in English is related to the grammatical category of definiteness. English has a strong tendency to avoid using bare nouns, i.e. nominal groups (especially in singular) should be mostly specified by either an article or another determiner. Possessive pronouns in cases such as *their* in Example 1 express definiteness even more explicitly than the definite article does, giving a monosemantic reference to the possessor. As a Slavic language without grammatical category of definiteness, Czech does not have such a strong tendency to express it. If our suggestion is correct, the situation in Russian should be more similar to Czech than to English, as it is also a language without obligatory expression of definiteness.

The translation of the sentence (4) into Russian supports this assumption: The use of the reflexive possessive pronoun *svoj* is grammatically correct but it is neither obligatory nor especially common.

(4)' RU: *В результате* (<своей>) *болезни они потеряли $ 1,8 млн заработной платы и других доходов.*

As shown in Table 2 for PCEDT-R, ca. 23% (39+15=54 occurrences) of English possessive pronouns are not expressed in Czech. For Russian, this number is lower: unexpressed pronouns make up ca. 15% (25+10=35 occurrences), but the difference with English is still statistically significant. In 13 cases (5%) in PCEDT-R, English possessive pronouns were not translated either into Czech or Russian. These are mostly the cases where the pronouns rather expresses definiteness than possessivity, cf. English *its* in *its first quarter* in Example 5.

(5) EN: *Bear Stearns reported improved earnings for <its> first quarter, ended Sept. 29.*—CZ: *Společnost Bear Stearns oznámila zvýšené výdělky za <> první čtvrtletí, končící 29. září.*—RU: *Bear Stearns сообщил об увеличившейся прибыли за <> первый квартал, закончившийся 29 сентября.*

There are occasional examples (one found in PCEDT-R, and some were found in Intercorp or can be constructed), where an English construction with a possessive was translated into Czech with a determiner, but with a different one than used in English. See Example 6 with the demonstrative pronoun *ten* in Czech. In Russian, the reflexive possessive *svoj* remains expressed (like in English).

(6) EN: *Lionel also urged holders of its stock and debt not to tender <their> securities...*—CZ: *Společnost Lionel též tlačí vlastníky svých akcií a dlužníky, aby <tyto> cenné papíry nenabízeli...*—RU: *Лайонел также убедил держателей его акций и долгов не номинировать <свои> ценные бумаги...*

Occasionally (4 instances in PCEDT-R), English possessives can also be translated into Czech or Russian with relative clauses (see Example 7 for Russian):

(7) EN: *Coupled with <his> current 1.2 million shares* [...] *the stake would have given him control of 55% of the concern.*—CZ: *Ve spojení s <jeho> současným 1.2 milionu akcií* [...] *by mu tento podíl poskytl kontrolu nad 55% podniku.*—RU: *Будучи соединена с 1,2 миллионами акций, которыми он в данный момент владеет* [...] *эта ставка обеспечила бы ему контроль 55% концерна.*

Observing the PCEDT-R data from the perspective of Czech and Russian, we can see that much fewer possessive expressions do not find their counterparts in English, than it was for the perspective English → Czech and Russian. For Czech (see Table 3), among 197 possessive and reflexive possessive pronouns, only 12 remain unexpressed (3 *no-poss* and 9 rewordings). Other cases are expressed either with possessive pronouns (174 cases), possessive nominal groups (e.g. *the company's* instead of *its*—5 cases) or the definite article (6 cases, see Example 8). For the perspective Russian → English & Czech (see Table 4), these numbers are even smaller.

(8) CZ: *Tento maloobchodník nebyl schopen najít pro <svoji> budovu kupce.*—EN: *The retailer was unable to find a buyer for <the> building.*—RU: *Компания была неспособна найти покупателя для <> здания.*

The interchangeability of a possessive pronoun and the definite article is especially interesting. It appears to be relatively systemic. Not only in PCEDT-R, but also in Intercorp, we can easily find examples where Czech and Russian possessive (or possessive reflexive) pronouns are aligned with the English definite article *the*. On the one hand it speaks in favor of our idea of correlation between the English grammatical category of definiteness and the frequency of possessive pronouns as compared to the Slavic languages Czech and Russian. On the other hand, it also means that possessive pronouns in Czech and Russian do not always express the possessivity exclusively. This meaning of reflexive possessive *svoj* and other possessive pronouns was described for Russian in Brykina (2009) but as far as we know no extensive research has been done for Czech.

The analysed examples allow us to confirm the proposed definiteness hypothesis, but more data need to be analysed to formulate the distributional rules more precisely. There are also some aspects of comparison that should be addressed in more detail in further analysis. For example, it seems that in cases where possessive pronouns fill actant positions in valency frames, they seem to be more frequent in Czech and Russian than in English (Example 9). Possessives in Czech and Russian tend to be more obligatory if the possessor's antecedent is more distant from the pronouns or belongs to a different clause (*судья Bork никогда не будет иметь высокого шанса на* ?*<свое>/*??*<его>/<>утверждение // судья Bork никогда не будет иметь высокого шанса на то, что <его>/*??*<0> утверждение будет одобрено*).

(9) CZ: *Byl mimo provoz od konce srpna, kdy mexická vláda oznámila <jeho> krach…*—EN: *It hasn't been operating since <it> was declared bankrupt by the Mexican government*—RU: *Он не действует с тех пор, как мексиканское правительство объявило о <его> банкротстве.*

6. Analysis: Czech ↔ Russian

6.1. Optionality

As we have observed in Section 4, Table 2 shows a significant difference in the frequency of possessive and reflexive possessive pronouns between Russian and Czech when translated from English possessive pronouns. It shows that possessives in Russian are used more frequently in translations. Why is it so? To answer this question, we have annotated optionality in PCEDT-R. Next to each sentence with a possessive or reflexive possessive pronoun, the label <OPT> was inserted in the cases where the possessive element could be omitted or inserted (if missing). The possibility to omit the pronoun does not necessarily infer that the meaning remains absolutely

the same, it is rather our assumption that in the same pragmatic context, the sentence could be also used without this pronoun, and the possessive meaning may be mostly reconstructed from the context.

Table 5. Optionality of possessive means in Czech and Russian

	expressed OPT	unexpressed OPT	total OPT	all nodes
cz refl_poss	27	13	50	238
cz poss	10			
ru refl_poss	36	2	54	
ru poss	16			

Table 5 shows a similar optionality in Czech and Russian:[8] Out of the translations of 238 English possessive pronouns, 50 and 54 cases, respectively (ca. 20% in both cases), are optional. However, in Czech, optionality was marked in a larger number of cases where possessivity was not expressed (13 cases in Czech vs. only two cases in Russian). Moreover, we observe a substantial difference in optionality of expressing possessivity between possessive and reflexive possessive in both languages: Reflexive possessives can be omitted more frequently (27 and 36 cases of reflexive possessives vs. 10 and 16 possessives in Czech and Russian respectively).[9] This fact appears especially interesting if we compare these numbers with the numbers in Table 2, showing that, in general, possessive pronouns are more frequently translated with possessive pronouns than with reflexive possessives in Czech and Russian.

A possessive pronoun is obligatory in cases where it fills a valency position of the inserted predication, if it expresses the contrastive meaning (e.g. *his magazines—many women's magazines* in Example 10) and so on.

(10) EN: *Today, Mr. Lang believes <his> magazines will offer what many women's magazines don't.*—CZ: *Dnes je Lang přesvědčen, že <jeho> časopisy nabízejí něco, co mnohé jiné ženské časopisy nemají.*—RU: *Сегодня г-н Лэнг считает, что <его> журналы предложат то, что не предлагают многие другие женские журналы.*

The possessive meaning is often lost when omitting the possessive expressive unit. In some cases it remains clear from the context, in other cases it does not, but still we suspect that there is no special need to express it. This makes our judgement about optionality rather weak and subjective. Nevertheless, we consider it to be very important as it helps us understand the graduality of this category in Slavic languages.

---

[8] This fact was not checked for statistical significance, because optionality is a subjective feature.

[9] The reasons for this difference can be rather empirical: reflexive possessive is coreferential to the subject, so possessor may be easier reconstructable from the context that in cases of non-reflexive possessivity. However, this has not been proved and will be eventually addressed in the future work.

Analysing examples with a different degree of optionality of possessive elements in Czech and Russian, we made some assumptions that can be confirmed or refused with a larger set of data and a more thorough analysis. For example, pronouns seem to be less obligatory when defining an inalienable part of the possessum. Also, in comitative constructions (A with B) a pronoun can be dropped out easier than in coordinative ones (A and B) as in Example 11.

(11) EN: *The play concerns Teddy 's homecoming with <his> wife of six years, Ruth.*—CS: *Hra se soustředí na Teddyho návrat domů s manželkou Ruth, se kterou je již šest let.*—RU: *Пьеса повествует о возвращении домой Тэдди и <его> жены Рут, на которой он женился за 6 лет до того.*

6.2. External possession

External possession (Haspelmath, 1999) is a phenomenon where a nominal unit is syntactically encoded as a verbal dependent but semantically understood as the possessor of one of its co-arguments. Krivan (2007) claims for Czech that the variability of semantic and syntactic properties of external possession constructions is higher than in other languages of the European linguistic area. We suppose that it is also higher in Czech than in Russian. Moreover, in Czech, external possession is expressed more frequently and specific steps towards grammaticalization of this phenomenon can be observed. In Czech, possessivity is often expressed by the Dative possessor *si*, which occurred in our examples parallel to English possessive pronouns, cf. Example 12.

(12) CZ: *Sběratelé, kteří <si> vydělali peníze na Wall Street, se stali více a více důležitou součástí obchodu s uměním*—EN: *Collectors who have made <their> money on Wall Street have become an increasingly important part of the art business.*

Moreover, in Czech, there is a gradual shift between the Dative possessor *si* and the so-called free Dative that does not contain the possessive meaning anymore (see Example 13 of colloquial Czech, where three datives are used at the same time). Free Dative is quite frequent in (especially colloquial) Czech, but the borderline between possessive and non possessive meaning is not clear in many cases.

(13) CZ: *Pustila jsem dceru na hory a ona <ti> <si> <mi> tam zlomila nohu!* (Jandová 1993:62, Cit. from Krivan 2007)—[lit. EN: *I let the daughter go to the mountains, and she <to you> <her> <to me> broke the leg there*].

In Russian, cases where external possession is expressed with a reflexive pronoun are marginal in our data and it may be used in cases where it is supported by the valency frame:

(14) RU: *Жители прокладывали <себе> путь через посыпанные стеклом улицы.*—EN: *Residents picked <their> way through glass-strewn streets.*—CZ: *Obyvatelé města <si> razili cestu ulicemi zasypanými sklem.*

It appears to be reasonable to address this topic in more detail. For example, it makes sense to compare the frequency of Czech *si* with the possessive meaning in translated and original texts. We assume that in original texts the frequency will be higher, because the meaning of this particle is synthesized with more difficulties when translating from English.

As for Russian, the question of external possessivity has been analysed e.g. in Haspelmath (1999), Kibrik (2003) and Brykina (2005), but from a different perspective. Addressing external possessivity in Russian on the corpus data in comparison with Czech and English is part of our plans for the future.

7. Conclusion

We have presented a corpus-based analysis of the use of possessive and reflexive possessive pronouns in the Prague Czech-English-Russian (PCEDT-R) parallel corpus. We have calculated the statistics of correspondences and analysed some tendencies that these statistics exhibit.

The created parallel data let us address differences in the expression of possessivity in the analysed languages more precisely. The statistics of pronoun correspondences in English, Czech and Russian and the interchangeability of English possessive pronouns with the definite article *the* proved the hypothesis of an existing correlation between the category of definiteness and the use of possessive pronouns. Furthermore, we analysed the differences between the use of possessive and reflexive possessive pronouns in Czech and Russian.

We believe that our findings may be interesting both from the theoretical and computational perspectives. From the perspective of computational linguistics, searching for rules of expressing possessivity helps us find and verify specific features in text that can be further used as background knowledge for the development of a multilingual tool for coreference and anaphora resolution; also, for machine translation, it is important to know which possessive pronouns should or should not be translated into Czech and Russian. From the theoretical point of view, our research contributes to contrastive comparative analysis of typologically related (Czech and Russian) and more distant (English vs. Czech & Russian) languages. The knowledge acquired by such comparison not only gives us the typologically relevant information in general but also an opportunity to know more about each separate language. For example, by comparing the specificity of use of possessive pronouns in Czech with Russian, we can understand more about each of these languages.

8. Acknowledgement

We acknowledge the support from the Grant Agency of the Czech Republic (grant 16-05394S). The work on this project was (partially) supported by the grant “Multilingual Corpus Annotation as a Support for Language Technologies” (LH14011, Ministry of Education, Youth and Sports). This work uses language resources developed,

stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References

1. *Bejček E., Hajičová E., Hajič J., Jínová P., Kettnerová V., Kolářová V., Mikulová M., Mírovský J., Nedoluzhko A., Panevová J., Poláková L., Ševčíková M., Štěpánek J., Zikánová Š.* (2013), Prague Dependency Treebank 3.0. Data/software, Charles university in Prague, MFF, ÚFAL, Prague, Czech Republic, http://ufal.mff.cuni.cz/pdt3.0/
2. *Boguslavsky I. M., Grigorieva S. A., Grigoriev N. V., Kreidlin L. G., Frid N. E.* (2000). Dependency Treebank for Russian: Concepts, Tools, Types of Information. Proceedings of the 18th Conference on Computational Linguistics. Vol 2, Saarbrücken, pp. 987–991.
3. *Brykina, M.* (2005), Types of constructions with an external possessor in Russian. [Tipy konstrukcij s vneshnim posessorom v russkom jazyke.] Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2005" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2005"].
4. *Brykina, M.* (2009), Coding the possessivity (corpus-based study of Russian) [Jazykovyje sposoby kodirovanija possessivnosti (na materiale korpusnogo issledovanija russkogo jazyka]. Ph.D. thesis, Moscow.
5. *Čmejrková S.* (1998), Syntactic and Discourse Aspects of Reflexivization in Czech. In: Hajičová, E. (ed.): Issues of Valency and Meaning, Prague, pp. 75–37.
6. *Daneš F., Hausenblas K.* (1962), Personal and reflexive possessive pronouns in Czech [Přivlastňovací zájmena osobní a zvratná ve spisovné češtině], Slavica Pragensia 4, 191–202.
7. *Dočekal M.* (2000), Reflexive possessives in Czech linguistics [Posesivní reflexivum v bohemistice], Studia minora facultatis philosophicae universitatis brunensis, Brno, pp. 47–59.
8. *Fed'ko E.* (2007), Non-canonical control of the reflexive pronoun in Russian [Nekanonicheskij kontrol refleksivnogo mestoimenija v russkom jazyke], Master Thesis, MSU, Moscow.
9. *Gebauer J.* (1890), Czech grammar for schools and teaching institutes [Mluvnice česká pro Školy střední a ústavy učitelské], Prague.
10. *Hajič J., Hajičová E., Panevová J., Sgall P., Bojar O., Cinková S., Fučíková E., Mikulová M., Pajas P., Popelka J., Semecký J., Šindlerová J., Štěpánek J., Toman J., Urešová Z., Žabokrtský Z.* (2012), Announcing Prague Czech-English Dependency Treebank 2.0. Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), European Language Resources Association, İstanbul, pp. 3153–3160.
11. *Haspelmath M.* (1999, External Possession in a European Areal Perspective. ln: Payne, D. L., Barshi, I. (eds.) External Possession, Amsterdam: Benjamins, pp. 109–135.

12. *Kibrik A. E.* (2003) External possession in Russian [Vneshnij possessor v russkom jazyke]. In: Kibrik A. E. Constants and variables in Russian [Konstanty i peremennyje jazyka], Sankt Peterburg, pp. 307–319.
13. *Křivan J.* (2007), External possession in Czech in typological and areal perspective [Externí posesivita v češtině: v typologické a areální perspektivě], Master›s thesis. Prague, Charles University in Prague.
14. *Marcus M., Santorini B., Marcinkiewicz M.-A., Taylor A.* (1999), Treebank-3LDC99T42. Philadelphia: Linguistic Data Consortium. Available online: https://catalog.ldc.upenn.edu/LDC99T42.
15. *Novák M., Nedoluzhko A.* (2015), Correspondences between Czech and English Coreferential Expressions, Discours: Revue de linguistique, psycholinguistique et informatique., Vol. 16, Caen, France, pp. 1–41.
16. *Padučeva E.* (1983), Reflexive pronoun with indirect antecedent and the semantics of reflexivity [Vozvratnoje mestoimenije s kosvennym antecedentom i semantika refleksivnosti], Semiotika i informatika, nr. 21, pp. 3–33.
17. *Padučeva E.* (1985). The statement and its relation to reality (referential mechanisms of the semantics of pronouns) [Vyskazyvanije i jego sootnesennost s dejstvitelnostju (referencialnyje aspekty semantiki mestoimenij)], Moscow.
18. *Panevová J.* (1986), About reflexive pronominalisation in Russian [K voprosu o refleksivnoj pronominalizacii v češskom jazyke], Linguistische Arbeitsberichte 54/56, pp. 44–56.
19. *Svoboda K.* (1880), Debate on the use of personal, possessive and reflexive pronouns in shortened compound sentences [Rozprava o užívání osobných, přisvojovacích a zvratných náměstek v souvětích zkrácených], Časopis českého muzea 54, pp. 124–142.
20. *Testelec J., Toldova S.* (1998), Reflexive pronouns in Dagestan languages and the typology of the reflexive [Reflexivnyje mestoimenija v dagestanskix jazykax i tipologija reflexiva], Тестелец Я. Г., Толдова С. Ю. (1998) Рефлексивные местоимения в дагестанских языках и типология рефлексива. // Questions of linguistics [Voprosy jazykoznanija], No4, pp. 35–57.
21. *Trávníček R.* (1951), Grammar of Czech [Mluvnice spisovné češtiny], Prague.

Data sources

1. *Czech National Corpus*—Intercorp [Český národní korpus—InterCorp]. Ústav Českého národního korpusu FF UK, Prague. Cit.02.02.2016 , accessible from WWW: <http://www.korpus.cz>.
2. *Prague Czech-English Dependency Treebank* (PCEDT)—Hajič J., Hajičová E., Panevová J., Sgall P., Bojar O., Cinková S., Fučíková E., Mikulová M., Pajas P., Popelka J., Semecký J., Šindlerová J., Štěpánek J., Toman J., Urešová Z., Žabokrtský Z. (2012), Announcing Prague Czech-English Dependency Treebank 2.0. Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), European Language Resources Association, İstanbul, pp. 3153–3160.

3. *Prague Dependency Treebank* (PDT)—Bejček E., Hajičová E., Hajič J., Jínová P., Kettnerová V., Kolářová V., Mikulová M., Mírovský J., Nedoluzhko A., Panevová J., Poláková L., Ševčíková M., Štěpánek J., Zikánová Š. (2013), Prague Dependency Treebank 3.0. Data/software, Charles university in Prague, MFF, ÚFAL, Prague, Czech Republic, http://ufal.mff.cuni.cz/pdt3.0/
4. *Prague Czech-English-Russian Dependency Treebank* (PCEDT-R)—Novák M., Nedoluzhko A., Schwarz (Khoroshkina) A. (2016): *planned to be published in Lindat/Clarin before the camera-ready version of this paper.*
5. *Russian Treebank* (RTB). Accessible from WWW: otipl.philol.msu.ru/~soiza/rtb/

RUSSIAN MINORITY LANGUAGES ON THE WEB: DESCRIPTIVE STATISTICS

Orekhov B. (nevmenandr@gmail.com),
Krylova I. (krylova93@gmail.com),
Popov I. (imvanya@gmail.com),
Stepanova E. (stepanovayekaterina@gmail.com),
Zaydelman L. (luda.zaidelman@yandex.ru)

National Research University Higher School of Economics, Moscow, Russia

The paper presents quantitative data about the web segments in minority languages of Russia. An ad-hoc search procedure allows to locate sites and pages on social networks that contain texts in a certain language of Russia. According to our data, there are texts in at least 48 of the examined languages on the Internet. We compared the gathered statistical data with the data from Wikipedia and the number of native speakers and found out that none of the "live" online data has a good correlation with the offline-life of language.

Keywords: minority languages, web as a resource, social networks, sociolinguistics

Acknowledgements: we thank Kirill Reshetnikov and Daria Ignatenko for collecting lexical markers for some of the languages and Dmitry Granovsky for his assistance in proofreading the text.

МИНОРИТАРНЫЕ ЯЗЫКИ РОССИИ В ИНТЕРНЕТЕ: ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Орехов Б. (nevmenandr@gmail.com),
Зайдельман Л. (luda.zaidelman@yandex.ru),
Крылова И. (krylova93@gmail.com),
Попов И. (imvanya@gmail.com),
Степанова Е. (stepanovayekaterina@gmail.com)

Национальный исследовательский университет Высшая школа экономики, Москва, Россия

В работе представлены количественные данные об интернет-сегментах на миноритарных языках России. Специальная технология поиска позволяет находить сайты и страницы в социальных сетях, на которых присутствуют тексты на одном из языков России. По нашим данным, в интернете присутствуют тексты по крайней мере на 48 языках из тех, которые мы обследовали. Мы сравнили собранные статистические данные с данными Википедии и числом носителей и обнаружили, что ни один параметр онлайн-данных не коррелирует с оффлайн-данными по языку (числом носителей).

Ключевые слова: миноритарные языки, малые языки, интернет как ресурс, социальные сети, социолингвистика

Благодарность: мы благодарим Кирилла Решетников и Дарью Игнатенко, которые собрали лексические маркеры для части языков, а также Дмитрия Грановского за помощь в работе над текстом.

1. Introduction

There are over a hundred national languages in Russia, excluding Russian and languages that are official in other countries. Some of these count more than a million native speakers. However, linguistic tools for all of these minority languages are equally scarce. The lack of tools—first and foremost, the corpora—results from the lack of digitized texts in said languages. The goal of our work is to form full collections of texts in Russian national languages, which can then be used to create text datasets (like annotated corpora, sets of n-grams and so on), and to count the number of sites where they are present.

We use several Internet sources to gather corpora—Yandex web-search as a word index store of the Internet, and the API of the most popular Russian social network Vkontakte to download texts from communities consisting of enthusiasts eager to speak a certain language.

2. Related works

Although papers dedicated to minority languages show a wide variety of approaches taken to the topic, they can generally be classified as quantitative vs. non-quantitative. Paper [4] presents a quantitative analysis of the Bashkir Internet, while such papers as [7], [8] display mostly qualitative analysis of Udmurt. There are also papers somewhere in between with little manual quantitative analysis [6].

There is an evident interest in investigating the Internet of minority languages, collecting corpora and developing NLP tools for such languages. In 2007 a huge project was launched by Scannell [9] for gathering texts of many minority and under-resourced languages. Even though are no particular corpora available, the website provides a lot of useful information about more than 2000 minority languages, including word n-grams and lists of urls to find texts on them.

Another topic worth mentioning is the contribution of modern technology to the well-being of minority languages. Mass media [12] argue that the Internet grants a new life to minority languages, anticipating an opportunity for the native speakers to talk to each other despite the distance. NLP researchers can also benefit from this as they gain access to new sources of texts in minority languages.

3. Methods

We use an almost automatic method for gathering corpora widely known as "seed words" and described in Wacky-papers [1], [2]. The method consists of seven stages, the first four of which were described in detail in [10]:

1. Search for lexical markers in grammars and phrasebooks.
   - Lexical markers are words that uniquely define the language that they belong to.

2. Search with Yandex.XML [15] for domains (websites) that contain the words obtained in stage 1.
   - We expect that texts found on pages in these domains are in the language to which the lexical marker belongs.

3. In the domains found in stage 2 find all pages that contain the lexical markers from stage 1 (send queries like "site: example.com 'marker'" to the Yandex search engine).

4. For each domain from stage 2 count the number of pages found in it on stage 3 and sort the domains into four groups: 1) download the whole domain, 2) download certain pages, 3) pages with files, 4) pages of social networks.
   - At the moment, we do not work with the third domain group.

5. Remove ambiguous domains from further processing (domains that were found several times using markers from different languages).

6. Download texts from the Vkontakte social network via its API [14], download texts from the Internet using a web-crawler (scrapy [13], Beautiful soup [11]).

7. For each text identify its language using the letter n-gram method.

Our methods have certain limitations. First of all, it is quite difficult to find a graphically unique word for a language. We gather and check lexical markers manually by running a search query that consists of a marker word and checking whether all found pages are in the expected language. However, considering that for some languages it might be very difficult to check every found page, there is no full confidence that several pages from the search result will not match another language.

Apart from that, we have query limits from the Yandex search engine—we can send only 1000 queries per day. As a result, the third stage of our method may take about 10–20 days for a widely spoken language with many lexical markers. The Yakut language with 450 140 natives and 22 lexical markers has the longest search history of 48 days. Some languages, on the other hand, can be searched in less than a day. The median search time over all 49 processed languages is 1,5 days.

Having collected all the urls, we also need to clean the lists and exclude, using a semi-automatic procedure, various music and video sites that contain only titles in the minority language. We also exclude the sites that were found by markers of different languages to consider them later.

When we download texts from a certain domain or a community, we already know the expected language of the obtained texts. However, the majority of these texts, especially when it comes to the texts from Vkontakte, are in fact in Russian or are useless and contain only links or pictures (see Fig. 2). Since our primary focus is compiling a collection of texts, we are only interested in texts in our target languages, so it is important for us to be able to identify the language of a text.

A common approach to identifying the language of a text is the letter n-gram method. The main idea of the method is to compare the list of the most frequent n-grams for the given text with the n-gram standards for the expected languages. The result of each comparison is a value called "distance" that represents the difference between the language of the text and the language with which it was compared. Once all the distances are calculated, the least one is determined and the corresponding language is considered to be the language of the text [3]. This method has proved to be very effective and quite accurate. However, the downside of the n-gram method is that in order to get the initial language standards it requires corpora, which, ironically, are exactly what we do not have.

Nevertheless, with a few modifications it is possible to apply the letter n-gram method to our problem. First of all, instead of creating n-gram standards for all languages, we only create a n-gram standard for Russian (more specifically, trigrams, which proved to provide the best results). Why do we assume that using a single standard would be enough? As already mentioned, we always know the expected language of our texts since they are downloaded from communities and domains already assigned to certain languages. As a result, there are only three options for each text: it may be written in the expected language, it may be written in Russian or it may be a "garbage" text, i.e. a text that consists entirely of Latin symbols (hyperlinks) or emoticons. Garbage texts can be easily identified and discarded by checking whether they contain cyrillic symbols. If a text contains cyrillic symbols and, therefore, is not garbage, we create its letter trigram standard and calculate the distance between Russian and the language of the text based on their standards. Finally, the distance value is compared with the empirically obtained threshold value, and in case the distance value is greater than the threshold, we can say that the text is not written in Russian or, in other words, is written in the expected language.

Unfortunately, it is very common for texts found on social networks to be quite short (like "спасибо" or "удачи"). These texts often cannot be correctly identified as Russian using just the letter n-gram method due to the skewed frequency data.

Nevertheless, we were able to improve language identification by additionally comparing texts against the list of Russian word unigrams. The modified method proved to be more effective when dealing with short texts.

4. Results

We are aware of 96 different minor languages that exist in the Russian Federation. So far, we have found lexical markers and searched for 49 of them, and have got some preliminary results for about 40 of them.

4.1. Social networks

For 30 minor languages we were able to find and download at least one Vkontakte community. A total of 1,735 communities were downloaded with 1,633 of them containing at least one text, where by text we mean either a post on the community wall or a comment under a post. Fig. 1 shows the distribution of communities by language.

In Fig. 2 you can see the distribution of texts by language. Similarly to the previous figure, bar heights represent the total number of posts extracted from communities that "speak" a corresponding language. A base-10 logarithmic scale is used to account for an order-of-magnitude difference between the figures for some of the languages; the black section of each bar represents the fraction of the posts that are actually written in this language. Clearly, for all languages the majority of texts are not written in said languages, but, instead, are either in Russian or are "garbage" texts.

A question then arises of whether there is any functional relationship between the total number of texts in a community and the number of texts that are actually written in the language to which the community is assigned. Applying linear regression to the data reveals that this might be the case: with the linear regression coefficient at 0.383 (p-value $< 2.2e^{-16}$), we can say that the number of texts that are written in the target language is 2.5 times less than the total number of texts in the community (see fig. 3).

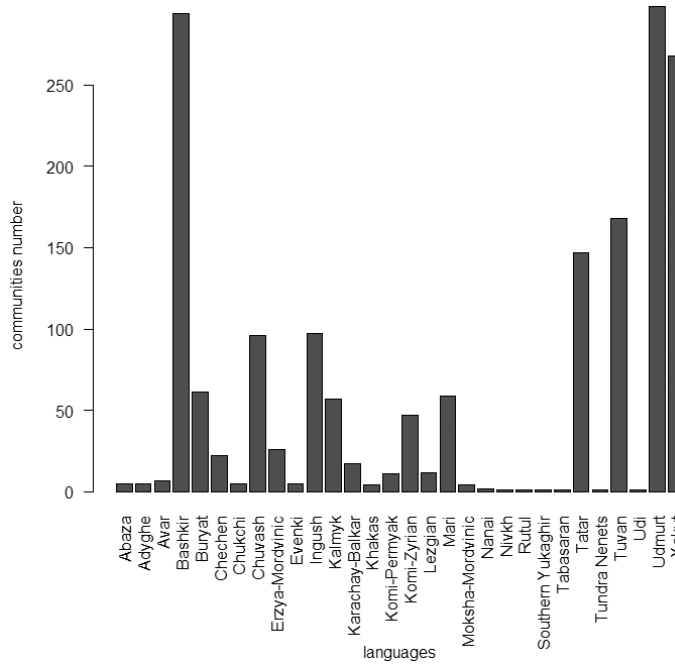


Fig. 1. Communities distribution by language

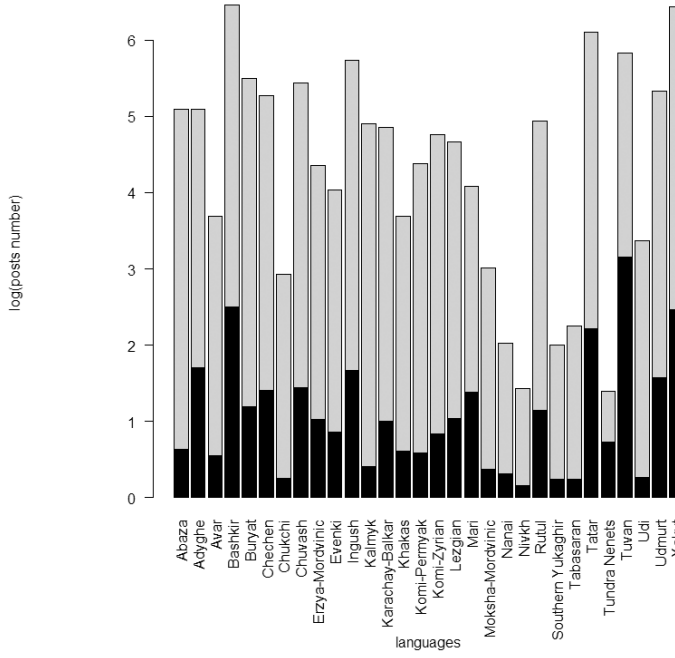


Fig. 2. Texts distribution by language

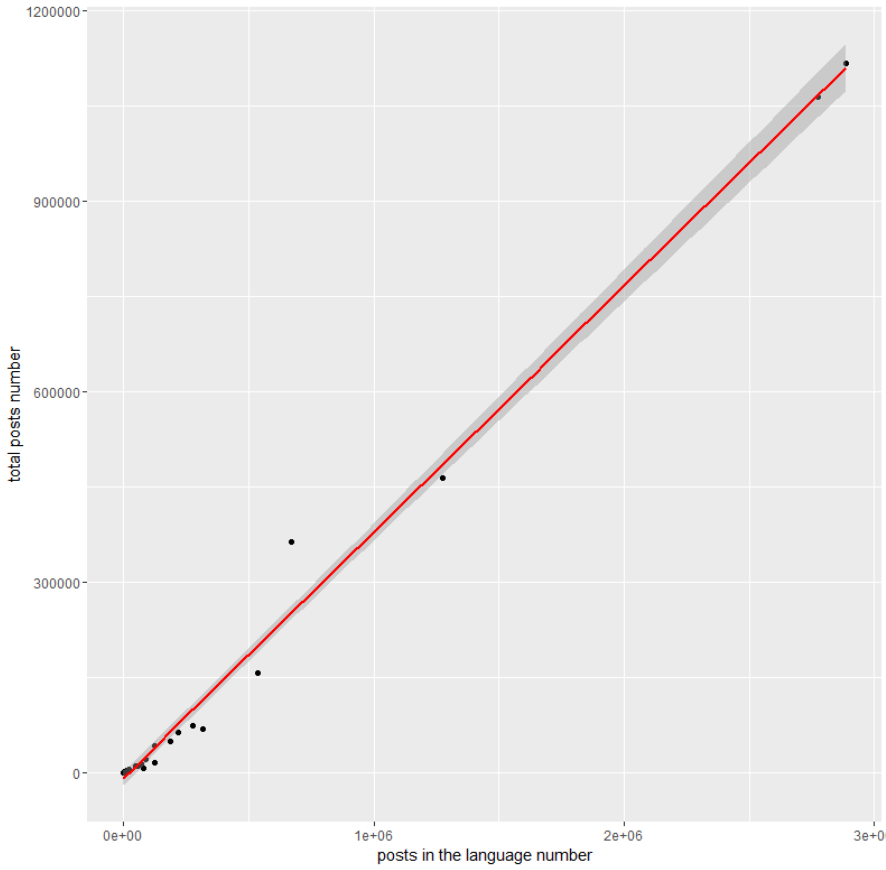


Fig. 3. Linear regression plot of total number of texts in communities attributed to a certain language vs. number of texts in said language

4.2. Internet

Currently, we have clean lists of urls (*example.com/page*) and domains (*example.com*) gathered for 49 minor languages. By clean lists we mean lists without non-informative sites such as music and video sites where a minor language might only be present in titles. It took us 239 search days to find 379 different "download-whole" language domains. We pay great attention to these domains, because, in our opinion, the number of such domains reflects language's self-sufficiency and its online development level.

In Table 1 you can see domain information for several groups of languages: the most active ones (Bashkir, Tatar and Yakut), the middle ones (Chuvash, Buryat, Kalmyk) and the least represented on the Internet (Rutul, Shor, Moksha-Mordvinic, Itelmen).

Table 1. Minor language domains information: ISO code, population, lexical markers found, domains found and days for search

ISO-639	Language	Natives	Markers	Domains	Days
ba	Bashkir	1,150,000	8	74	19
tt	Tatar	4,280,000	3	59	12
sah	Yakut	450,140	22	52	48
cv	Chuvash	1,152,404	5	20	8
bxr	Buryat	283,000	9	19	9
xal	Kalmyk	80,546	13	5	6
rut	Rutul	30,360	6	1	0.5
cjs	Shor	2,839	2	0	0.2
mdf	Moksha-Mordvinic	2,025	8	0	3.5
itl	Itelmen	7	2	0	0.5

Below you can see the histogram for all the processed minor languages. Please note that 19 languages with zero domains were excluded from the list for readability purpose. The exclude languages are: Aleut, Archi, Even, Forest Yukaghir, Gorno-Altai, Itelmen, Koryak, Kubachi-Ashtin, Mansi, Moksha-Mordvinic, Nanai, Nivkh, Nogai, Shor, Tabasaran, Tofa, Tsakhur, Tundra Yukaghir, Udi.

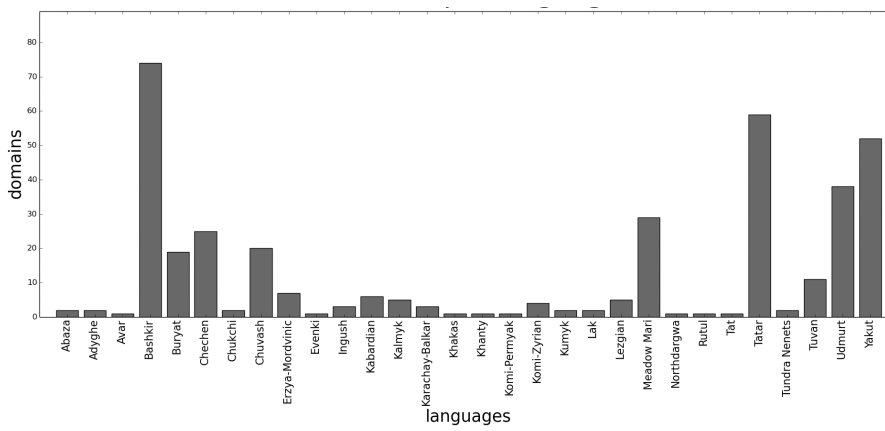


Fig. 4. Domains distribution by language

4.3. Comparison with Wikipedia data

When you set out to collect texts for a corpora, Wikipedia seems to be an obvious choice as a source for texts. Indeed, Wikipedia is basically a very large collection of texts, and many languages of Russia have their own language-specific Wikipedia. However, as was shown in [5], Wikipedia can sometimes be an inadequate representation of a language.

With actual data at our disposal, we were able to compare languages based on their parameters derived from the data. We created a dataset where rows corresponded to the languages and columns to the following parameters:

- the number of articles on the Wikipedia in the given language in 2014
- the number of articles on the Wikipedia in the given language in 2015
- the number of communities on Vkontakte in the given language
- the number of texts in this language in Vkontakte communities
- the number of tokens in this language in Vkontakte texts
- the number of whole-download domains for this language
- the number of webpages (urls) with at least a word in the minor language
- the number of tokens in this language on downloaded webpages
- the number of native speakers

We used this dataset to check if there were correlations between the parameters. We removed all the languages for which we do not have full information, e.g. for Bashkir we have not downloaded or processed texts from the web, and for Kumyk we have not processed Vkontakte communities. With such languages removed, there are 33 languages for which we have data both from Vkontakte and the web. We used this data as input to calculate linear correlations, the results are presented in Table 2, in which each cell contains a correlation coefficient for the corresponding parameters, and cells with values greater than 0.7 are grayed out.

Some of the correlations are rather obvious, e.g. the number of Wikipedia articles in 2014 vs. 2015 or the number of Vkontakte communities that "speak" a certain language vs. the number of texts in this language on Vkontakte, number of tokens on the web vs number of urls. However, the table also reveals something less trivial: neither the number of Vkontakte communities using a language nor the number of domains for this language correlates with the number of the language's native speakers. This means that the relationship between an actual language and its online representation is at least nonlinear, and that a language is used differently online.

Table 2. Pearson correlation coefficients for language parameters

	wiki 2014	wiki 2015	communities	posts in language	tokens on vk	domains number	urls	tokens in web	native speakers
wiki 2014	1	0.978	0.263	0.008	0.024	0.469	0.674	0.612	0.465
wiki 2015	0.978	1	0.376	0.131	0.142	0.592	0.668	0.660	0.617
communities	0.263	0.376	1	0.925	0.895	0.825	0.380	0.525	0.401
posts in language	0.008	0.131	0.925	1	0.982	0.714	0.064	0.346	0.299
tokens on vk	0.024	0.142	0.895	0.982	1	0.726	0.068	0.357	0.270
domains number	0.469	0.592	0.825	0.714	0.726	1	0.540	0.684	0.587
urls	0.674	0.668	0.380	0.064	0.068	0.540	1	0.765	0.270
tokens in web	0.612	0.660	0.525	0.346	0.357	0.684	0.765	1	0.470
native speakers	0.465	0.617	0.401	0.299	0.270	0.587	0.270	0.470	1

5. Conclusion and future plans

So far, we have briefly analyzed about 40 minor languages. We found out the most active languages in the Vkontakte social network, as well as the ones which are most represented on the Internet in general (based on the number of domains written predominantly in the minor language). We compared the gathered statistical data with the data from Wikipedia and the number of native speakers and found out that none of the "live" online data has a good correlation with the offline-life of language.

That means that, according to our data, the offline life of a language is completely different from its online existence. Obviously, representation of the language on the Web, affected by various factors, is not limited to the number of its speakers. We have found possible indicators of language activity on the Internet—the number of webpages and the number of communities on social networks, which demonstrate the degree of the web-vitality of a language. However, these assumptions require further research.

We plan to continue and deepen our analysis. As of today, we have already downloaded texts from all the domains and communities for 41 languages, tagged the downloaded texts with the language identification tool and counted the number of distinct webpages and posts per language, assessed the amount of minority-language text in them and counted token information for them. We plan to perform thematic analysis of websites and Vkontakte communities and compare our experience and results with the works of foreign minor language researchers and with those of social linguistics researchers.

All text collections and additional data are available on our website http://web-corpora.net/minorlangs/ (as of today, the website only has a Russian interface).

References

1. *Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.* (2009), The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation, Vol. 43(3), pp. 209–226.
2. *Boleda, G., Bott, S., Meza, R., Castillo, C., Badia, T., López, V.* (2006), CUCWeb: a Catalan corpus built from the Web, Proceedings of the 2nd International Workshop on Web as Corpus, Trento, pp. 19–26.
3. *Cavnar, W. B., Trenkle, J. M.* (1994), N-gram-based text categorization. Ann Arbor MI, Vol. 48113(2), pp. 161–175.
4. *Orekhov B. V., Gallyamov A. A.* (2013), Bashkir internet: lexis and pragmatics in a quantitative aspect [Bashkirskiy internet: leksika i pragmatika v kolichestvennom aspekte], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2012"], Bekasovo, pp. 502–509.

5. *Orekhov B. V., Reshetnikov K. Yu.* (2014), To the assessment of Wikipedia as a linguistic source [K otsenke Vikipedii kak lingvisticheskogo istochnika], Contemporary Russian on the Internet [Sovremennyy russkiy yazyk v internete], Moscow, Languages of slavic culture [Jazyki slavjanskoy kul'tury], pp. 310–321.
6. *Pischlöger C.* (2014), Notes from Murjol Underground: super udmurts in cyberspace [Zapis(k)i iz Murzhol Undeground: Super udmurty v Cyberspace], Proceedings of IV international science-practical conference "Florov's readings" [Trudy IV Mezhdunarodnoy nauchno-prakticheskoy konferentsii "Frolovskie chteniya"], Glazov, pp. 56–59.
7. *Sacharnych D. M.* (2006), National Udmurt Internet: what interferes with the development [Udmurtskiy natsional'nyy internet: chto meshaet razvitiyu?], available at: http://udmurt.info/pdf/texts/udmint.pdf
8. *Sakharnykh D. M.* (2008), New in national Udmurt Internet development: winter-autumn 2008 [Novoe v razvitii udmurtskogo natsional'nogo interneta: Zima-osen' 2008], available at: http://udmurt.info/pdf/texts/udmurtnet-zima-osenj-2008.pdf
9. *Scannell, K. P.* (2007), The Crúbadán Project: Corpus building for under-resourced languages. Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop, Vol. 4, pp. 5–15.
10. *Zaydelman L., Krylova I., Orekhov B., Stepanova E.* (2015), Languages of Russia: Using Social Networks to Collect Texts, (in print) In Proceedings of the 9th Summer School in Information Retrieval and Young Scientist Conference (RuSSIR 2015)—Revised and Selected Papers, Communications in Computer and Information Science, St. Petersburg, Vol. XXX, pp. 1–8.
11. *Beautiful Soup*—http://www.crummy.com/software/BeautifulSoup/bs4/doc/
12. New technologies contribute to the preservation of minor languages [Novye tekhnologii sposobstvuyut sokhraneniyu malykh yazykov]—http://www.polit.ru/news/2014/01/15/ps_lang/
13. *Scrapy*—http://doc.scrapy.org/en/latest/
14. *VK API*—https://vk.com/dev/api_requests
15. *Yandex.XML*—https://tech.yandex.ru/xml/

К СЕМАНТИКЕ РУССКОГО ВИДА: МОМЕНТ НАБЛЮДЕНИЯ И ДИСКУРСИВНЫЙ КОНТЕКСТ[1]

Падучева Е. В. (elena.paducheva@yandex.ru)

ФИЦ ИУ РАН Москва, Россия

Во **внутрифразовом** контексте грамматический вид (если отвлечься от его функции выражения многократности) характеризует ситуацию по отношению к **моменту наблюдения**. Момент наблюдения хорошо работает также в **дискурсивном** контексте, если речь идет о последовательности из двух одинаковых видов. Так, понятие момент наблюдения позволяет вывести текстовые (дискурсивные) значения форм СВ и НСВ из их значений в составе изолированного высказывания. Возникает вопрос, что происходит при соположении или сочинении разных видовых форм. Предметом внимания в докладе является морфосинтаксическая конфигурация «глагол СВ + союз И + глагол НСВ». Показано, что при установлении временных соотношений между перфективной и имперфективной ситуацией в составе указанной конфигурации момент наблюдения тоже играет важную роль. Если момент наблюдения глагола НСВ приходится на перфектное состояние глагола СВ, возникает **секвентное** отношение между ситуациями — имперфективная ситуация после перфективной; если на само событие, то имперфективная и перфективная ситуация **одновременны**.

Ключевые слова: дискурсивный контекст, момент наблюдения, секвентность, одновременность

TOWARDS SEMANTICS OF THE RUSSIAN ASPECT: MOMENT OF PERCEPTION AND DISCURSIVE CONTEXT

Paducheva E. V. (elena.paducheva@yandex.ru)

Federal Research Center Informatics and Management, Moscow, Russia

In the **context of a sentence** grammatical aspect (apart from its function of expressing multiplicity y) characterizes a situation with respect to the **moment of perception**. In the **context of discourse** the moment of perception also honestly does its job in the case of a sequence of identical aspectual forms. In fact, the notion “moment of perception” makes it possible

[1] Данная работа выполнена при финансовой поддержке РГНФ, грант № 14-04-00604а

to derive the textual (discursive) meaning of the Perfective and Imperfective forms from their meaning in an isolated utterance. The question arises: what happens in the context of juxtaposed or conjoined different aspectual forms. The subject of attention in the paper is a morphosyntactic configuration "Perfective verb + conjunction *i* 'and' + Imperfective verb". It is demonstrated that temporal relationships between situations in the context of this configuration heavily depend upon the moment of perception. If the perception moment of the Imperfective verb is synchronous to the perfective state of the Perfective one the **sequential** relationship between situations arises. If the perception moment of the Imperfective verb is synchronous to the event denoted by the Perfective verb the relationship between situations is **synchronicity**.

Key words: Perfective, Imperfective, moment of perception, context of discourse

Вид, если отвлечься от его функции выражения многократности, характеризует ситуацию по отношению к **моменту наблюдения** (в отличие от времени, которое, в своем основном значении, характеризует ситуацию по отношению к **моменту речи**, ср. немецкий термин Betrachtzeit, момент наблюдения, который используется по отношению к виду и противопоставлен термину Redezeit, момент речи, который используется по отношению к времени: Kratzer 1977, Bäuerle 1979 и др.).

Так, общим свойством глаголов совершенного вида (СВ) является то, что они задают **ретроспективный взгляд на ситуацию**, иначе — ретроспективный ракурс (retrospective **viewpoint** в терминах Smith 1997; ретроспективную перспективу; ретроспективную точку отсчета по Рейхенбаху, или ретроспективную позицию наблюдателя в терминах Падучева 1996: 12, 269). На ретроспективную позицию наблюдателя у глагола СВ указывают: отсутствие у глагола СВ формы наст. времени (наст. время порождает синхронного наблюдателя); несочетаемость с фазовыми глаголами, с показателями включенного времени и длительности.

А глагол несовершенного вида (НСВ) может выражать синхронный и ретроспективный ракурс в прошедшем времени; синхронный и проспективный в будущем (Падучева 2010). Но в настоящем времени, при речевом, т. е. не нарративном, режиме интерпретации, глагол НСВ выражает однозначно **синхронный** ракурс — синхронную точку отсчета. Синхронная точка отсчета была понята как одновременность моменту наблюдения (см. о моменте наблюдения Гловинская 1982: 128).

1. Дискурсивный контекст

Понятие момент наблюдения хорошо работает во внутрифразовом контексте. А также в дискурсивном контексте, если речь идет о последовательности из двух одинаковых видов. Так, понятие момент наблюдения позволяет вывести текстовые (дискурсивные) значения форм СВ и НСВ из их значений в составе изолированного высказывания. А именно.

1. Форма НСВ в актуально-длительном значении предполагает момент наблюдения, расположенный «в середине» ситуации (Падучева 1996: 363). Отсюда тот факт, что соположенные или сочиненные формы НСВ выражают **одновременные** процессы или состояния. Например.

(1.1) Компания охотников *ночевала* в мужицкой избе на свежем сене. В окна *глядела* луна, на улице грустно *пиликала* гармоника, сено *издавало* приторный, слегка возбуждающий запах. [А. П. Чехов. Зиночка]

2. Форма СВ обозначает событие и предполагает момент наблюдения, расположенный «в середине» перфектного (т. е. финального) состояния события, которое отстоит во времени от исходного состояния. Отсюда тот факт, что соположенные или сочиненные формы СВ выражают **последовательные** события. Например.

(1.2) Медвежонок *схватил* веник, *подмёл*. Ёжик *распахнул* дверь. Все трое *вошли* и *поздоровались*. [Сергей Козлов. Новогодняя сказка // «Мурзилка», 2003]

Примечание. Сочиненные глаголы СВ могут, в определенном контексте, обозначать одновременные события (см. об одновременности в Падучева 1996: 363). Так, в (а) моментальные события *вздрогнул* и *уронил*, скорее всего, одновременны:

(а) Первоклассник, стоявший рядом с её кабинетом, *вздрогнул и уронил* на пол деревянный пенал. [Андрей Геласимов. Фокс Малдер похож на свинью (2001)]

Пример (б) может заставить подумать, что возможен обратный порядок следования ситуаций: *видел* до того, как *вернулся*. Однако это, скорее всего, просто небрежность; имеется в виду не *вернулся и видел*, а *летал и видел*:

(б) И знаете, Константин Аркадьевич, он летал очень далеко, только что *вернулся и видел* нечто необычайное. [И. А. Ефремов. Бухта радужных струй (1944)]

Возникает вопрос, что происходит при соположении или сочинении разных видовых форм. Предметом внимания в докладе является морфосинтаксическая конфигурация «глагол СВ + союз И + глагол НСВ».

Известно отличие русского языка (и других восточнославянских) от западных славянских, в частности, от чешского, состоящее в том, что форма прошедшего времени несовершенного вида с **секвентной** функцией, которая в русском языке считается маргинальной, в чешском нормальна. В Петрухина 2000: 84, приводятся примеры (1.3а), (1.4а) (здесь даются буквальные переводы на русский чешских фраз), где формы несов. вида в чешском вполне нормативны:

(1.3) а. *Он схватил меня за руку и *тянул*.

(1.4) а. *Увидев, что все бесполезно, я замолчала и *молилась*.

По-русски вместо (1.3а) надо сказать (1.3б), вместо (1.4а)–(1.4б):

(1.3) б. Он схватил меня за руку и *стал тянуть* (или: и *потянул*).

(1.4) б. <…> я замолчала и *стала молиться*.

Т. е. в русских переводах надо заменить сочетание «глагол СВ + глагол НСВ» на два глагола СВ.

Однако если задать в Корпусе конфигурацию «глагол СВ + союз И + глагол НСВ», то выдается больше 13 тыс. примеров. В подкорпусе со снятой омонимией примерно 200 примеров. (Поиск по Корпусу был ограничен глаголами изъявительного наклонения и прош. времени. Кроме того, был исключен глагол *быть*, который имеет преимущественно служебные употребления.)

Так или иначе, ясно, что конфигурация «глагол СВ + союз И + глагол НСВ» в русском языке активно воспроизводится. Задача — выяснить, как она понимается. (Близкая задача была поставлена в Князев 2014.)

Форма НСВ в актуально-длительном значении предполагает момент наблюдения, расположенный «в середине» ситуации (Падучева 1996: 363). Именно так интерпретируется НСВ, например, в первой фразе примера (1.1). Значит, вопрос в том, чтобы понять, как обеспечивается такой момент наблюдения у глагола НСВ в составе нашей конфигурации.

Гипотеза. У русского глагола НСВ в норме нет секвентной функции; т. е. форма НСВ не выражает отношения **следования во времени** между событием, которое обозначается глаголом СВ, и процессом или состоянием которое обозначается глаголом НСВ. Имеет место **одновременность** процесса или состояния, обозначенного глаголом НСВ, с событием, обозначенным глаголом СВ. Точнее, часто, одновременность не с самим событием, а с его перфектным состоянием. Момент наблюдения (на событии или на перфектном состоянии) задается глаголом СВ и остается неизменным при переходе к следующему за ним глаголу НСВ.

Рассмотрим три возможности, которые соответствуют трем акциональным классам глагола НСВ — это может быть состояние, неагентивный процесс и деятельность (агентивный процесс).

2. Глагол НСВ выражает состояние

В примерах (2.1)–(2.7) момент наблюдения для глагола НСВ задается глаголом СВ или его перфектным состоянием. Ситуация, выраженная глаголом НСВ, связана отношением одновременности с событием, выраженным глаголом СВ, или, чаще, с перфектным состоянием этого события.

(2.1) К тому моменту мы уже *подружились и питали* друг к другу нежные чувства. [Сати Спивакова. Не всё (2002)]

(2.2) Я уже *разобрался и знал*, что всё это — бесчисленные и разноюродные тётки и бабки Людмилы и покойной Ани: [Андрей Волос. Недвижимость (2000) // «Новый Мир», 2001]

(2.3) Такое сравнение он *придумал и гордился* его художественностью. [И. Грекова. Фазан (1984)]

(2.4) Он почти *выздоровел и собирался* уйти к партизанам, когда немцы напали на его след. [Р. Б. Ахмедов. Промельки (2011) // «Бельские Просторы»]

(2.5) Потом дешёвую, яркую Нюра выпросила для себя: она на самом деле *заболела и думала*, что икона поможет. [Юрий Трифонов. Предварительные итоги (1970)]

(2.6) Он *поднялся и хотел* выйти, но тут быстро вошла мать с керосиновой лампой в руках. [Ю. О. Домбровский. Факультет ненужных вещей, часть 5 (1978)]

(2.7) Он *озверел и хотел* убить меня из пистолета. [Алексей Слаповский. Гибель гитариста (1994–1995)]

У глагола СВ *устать* фактически нет событийного значения — он обозначает только перфектное состояние. Так что в (2.8) одновременны не состояние и событие, а два состояния (то же верно для *озвереть* в (2.7)):

(2.8) повторю, я просто *устала и хотела* немного отдохнуть. [Андрей Митьков. Мороз по коже. //«Известия», 2003.01.08]

В примерах (2.1) и (2.4) момент наблюдения, общий для двух пропозиций, задан обстоятельством времени (*к тому моменту* в (2.1), *когда* … в (2.4)).

В (2.2) имеется причинно-следственное, а не только чисто временное отношение между перфективной и имперфективной ситуацией.

Состояния не имеют четкого начала, так что о начинательности в значении формы НСВ речи нет: имперфективная ситуация наблюдается изнутри, в своей срединной фазе.

Поскольку перфектное состояние идет после события, а момент наблюдения у глагола НСВ приходится на перфектное состояние события, выраженного глаголом СВ, возникает отношение **временной последовательности** между ситуацией, выраженной глаголом НСВ, и ситуацией, выраженной глаголом СВ, см. примеры (2.2), (2.3). А если момент наблюдения приходится на само событие, то между имперфективной и перфективной ситуацией отношение

одновременности, см. примеры (2.1), (2.4). Возможна одновременность двух состояний, перфектного и имперфективного, см. пример (2.8).

3. Глагол НСВ выражает неагентивный процесс

Примеры.

(3.1) Но оно <солнце> *взошло и подсвечивало* влажные валы тумана, а от ручья по низкому лугу тянулся запах таволги и хвоща. [Юрий Коваль. Гроза над картофельным полем (1974)]

Форма СВ *подсветило* выражала бы последующее событие, а НСВ *подсвечивало* выражает одновременность процесса перфектному состоянию события, обозначенному глаголом СВ — т. е. это процесс, наступающий после события (сначала взошло, потом подсвечивало). Можно выразить начинательность (*стало подсвечивать*), тогда ситуации будут представлены как последовательные во времени события.

(3.2) <…> мороз, который вдруг *грянул и затруднял* передвижения. [Юрий Трифонов. Дом на набережной (1976)]

В (3.2) глагол НСВ выражает постоянно действующий фактор (а не событие, как было бы при глаголе СВ *затруднил*). Тут одновременность самому событию, а не перфектному состоянию, и между ситуациями отношение одновременности.

(3.3) Она жестоко исхудала, грязно-бурая шерсть на ней *свалялась и висела* клочьями. [Юрий Коваль. У Кривой сосны (1979)]

В (3.3) тоже нет идеи следования во времени: глагол СВ *свалялась* имеет акцент на перфектном состоянии (как и *устал*). Имеет место одновременность двух состояний; имеется синхронный момент наблюдения, общий для перфектного состояния глагола СВ и стативной ситуации, обозначаемой глаголом НСВ *висела*.

(3.4) Путаница наростов и тяжей *ослабла и болталась* вокруг её бёдер, покрытых сплошь растяжками и ссадинами. [Людмила Улицкая. Казус Кукоцкого (Путешествие в седьмую сторону света) // «Новый Мир», 2000]

Можно выразить начинательность (*стала болтаться*), но это будет другой смысл. То же в (3.5) — *воняло* [1] *стало вонять*:

(3.5) хотя всё сгнило, *протухло и воняло*, [Анатолий Рыбаков. Тяжелый песок (1975–1977)].

(3.6) Рукав рубахи *оторвался и висел* на одной ниточке. [Ю. О. Домбровский. Обезьяна приходит за своим черепом, часть 3 (1943–1958)]

(3.7) Белые и красные пятна фонарей *приблизились и скакали* по земле. [Ю. О. Домбровский. Обезьяна приходит за своим черепом, часть 2 (1943–1958)]

(3.8) Глаза его *сузились и слезились* от порывов ветра. [Геннадий Башкуев. Маленькая война // «Сибирские огни», 2013]

Очевидно, и здесь момент наблюдения приходится на середину имперфективной ситуации и перфектного состояния перфективной. В (3.5)–(3.8) между ситуациями отношение одновременности — несмотря на то, что момент наблюдения приходится на перфектное состояние глагола СВ.

При этом форма НСВ не выражает начинательности.

Примечание. Начинательность составе имперфективной пропозиции может быть категорически невозможна. Во-первых, в том случае, если у глагола СВ акцент на перфектном состоянии и нет событийной семантики (как в *устать*), ср. *утомились* в (3.9).

(3.9) Волшебные чёрные кони и те *утомились и несли* своих всадников медленно, и неизбежная ночь стала их догонять. [М. А. Булгаков. Мастер и Маргарита, часть 2 (1929–1940)]

Во-вторых, и глагол НСВ играет роль — почему-то нельзя сказать **стал валяться*:

(3.10) роговой гребень из головы ее *выпал и валялся* у ног на полу, [Роман Шмараков. Камеристка кисти Клотара // «Сибирские огни», 2013]

В-третьих, начинательность неуместна, если глагол НСВ выражает состояние на сверхдолгом интервале, см. (3.11). Кроме того, *привыкнуть* (как и *устать*) — глагол с акцентом на перфектном состоянии; в этом смысле (3.11) похоже на (3.9):

(3.11) к нему *привыкли и переносили* его присутствие как неизбежное зло. [И. С. Тургенев. Певцы (1850)]

В итоге, во всех примерах момент наблюдения приходится на срединную фазу имперфективной ситуации.

4. Глагол НСВ выражает деятельность, т. е. агентивный процесс

Чаще всего, здесь, как и ранее, имперфективная ситуация одновременна перфектному состоянию события, и если момент наблюдения приходится на срединную фазу перфектного состояния, то имперфективная ситуация наступает после события. А если на событие, то имперфективная и перфективная ситуация одновременны.

(4.1) «Мы *обнялись и плакали* непонятно от чего», — вспоминал потом Володя. [Сати Спивакова. Не всё (2002)]

Если бы между ситуациями было отношение следования во времени, то нужно было бы выразить начинательность. Между тем (4.1) ≠ *обнялись и заплакали* и ≠ *обнялись и стали плакать*. Скорее всего, в тот момент, когда обнялись, уже плакали. Хотя предложение неоднозначно — его можно понять в значении следования во времени.

(4.2) — Как это нет?.. — он *замолчал и смотрел* на меня в полном изумлении. [Андрей Геласимов. Ты можешь (2001)] <если смотрел молча, значит отношение следования>

(4.3) С готовностью *застыл и дожидался* по стойке «смирно». [Олег Павлов. Карагандинские девятины, или Повесть последних дней // «Октябрь», 2001]

(4.4) Оба швейцара (один пил кофе, а другой— чай) напились, *оделись и ждали* девяти часов, когда приносят письма и газеты. [И. А. Гончаров. Май месяц в Петербурге (1891)]

В (4.5) момент наблюдения приходится на середину имперфективной ситуации. Момент, в который Петр Сергеевич сел у стола, опущен в описании происходящего. Отношение между ситуациями — временная последовательность:

(4.5) Петр Сергеевич уже *проснулся и сидел* у стола, почесываясь и перелистывая старый номер «Пути», который Андрей выменял вчера у цыгана на банку пива, но так и не стал читать. [Виктор Пелевин. Желтая стрела (1993)]

В (4.6) момент наблюдения приходится на срединную фазу имперфективной ситуации; но *стоял* фактически означает ‘продолжал стоять’, так что отношением между ситуациями не может быть следование:

(4.6) Месяца два назад, в мае, Денис Иванович прощался у калитки с гостями, *простился и стоял* у калитки, дыша утренней прохладой, [Алексей Слаповский. Гибель гитариста (1994–1995)]

В (4.7) можно было бы сказать *и стала знакомиться*, но это был бы другой смысл. NB *тогда*, которое задает момент наблюдения, общий для перфективной и имперфективной ситуации. Между событием и деятельностью было бы отношение временного следования — если бы не общее *тогда*, которое выражает одновременность.

(4.7) Она тогда только что *приехала и знакомилась* с нашим учебным музеем. [Ю. О. Домбровский. Факультет ненужных вещей, часть 1 (1978)]

В (4.8) *сразу* дает акцент на событии, перфектное состояние отсутствует, и между ситуациями отношение временного следования:

(4.8) Это гостеприимство по-восточному мы сразу *почувствовали и ощущали* постоянно: [И. К. Архипова. Музыка жизни (1996)]

Из-за *сразу* у глагола СВ нет перфектного состояния: *сразу* дает событийный акцент. Поэтому предикация *ощущали постоянно* имеет свой момент наблюдения.

В (4.9) между двумя ситуациями нет отношения следования во времени; тем более, что НСВ *заведовал* имеет акциональный класс обобщенная деятельность, которая локализуется в сверхдолгих интервалах. Это конъюнкция двух пропозиций; по умолчанию, отношение между ситуациями — одновременность.

(4.9) Ох ты господи, да Метальников, к вашему сведению, ещё до революции во Францию *уехал и заведовал* в Институте Пастера отделом. [Даниил Гранин. Зубр (1987)]

В (4.10) тоже нет следования во времени: ходила не ПОСЛЕ того, как *выздоровела*, а ПОТОМУ ЧТО *выздоровела*. Начинательность неуместна, *ходила ≠ стала ходить* — деятельность рассматривается сразу в срединной стадии своего развертывания, что нормально для узуального значения НСВ:

(4.10) Но постепенно чувство вины за письмо и радость от того, что я *выздоровела и ходила* в «нормальную» школу, начали притупляться. [Анатолий Алексин. Раздел имущества (1979)]

Аналогичен пример (4.11): пропозиции связаны не отношением временного следования, а причинно-следственным; глагол НСВ *сидела* здесь обозначает устойчивое состояние:

(4.11) Чего ты добиваешься? Чтобы я *разошлась и сидела* у тебя под юбкой? [Токарева Виктория. Своя правда // «Новый Мир», 2002].

В (4.12) тоже причинно-следственная, а не временная связь между перфективной и имперфективной ситуацией (разве что действует принцип «propter hoc ergo post hoc»):

(4.12) Станцию почти каждый день обстреливали тяжёлые немецкие орудия, — немцы наловчились, *пристрелялись и лепили* снаряды метко, по стенам цехов, грохот разрывов то и дело потрясал землю. [Василий Гроссман. Жизнь и судьба, ч. 1 (1960)]

Вернемся к случаю, когда временное соотношение релевантно. Хорошо чувствуют себя в контексте конфигурации «глагол СВ + союз И + глагол НСВ» глаголы, обозначающие неактивную деятельность: *сидеть, стоять, ждать*: момент наблюдения находится в срединной фазе имперфективной ситуации, см. примеры (4.1)–(4.7). Если же глагол НСВ обозначает активную деятельность, то момент наблюдения часто не может попасть в срединную фазу имперфективной ситуации. Чтобы возникло секвентное отношение между перфективной ситуацией и имперфективной, приходится выразить начинательность, которая дает глагол СВ и новый момент наблюдения. Так, в (4.13) лучше сказать *лёг и стал пить воду*; в (4.14) лучше сказать *вернулся и стал топать*, и т. д.

(4.13) Когда они вышли к Волге, парень *лёг и пил воду*, а напившись, бережно стряхнул на ладонь капли с ватника и слизал их, как голодный крошки хлеба. [Василий Гроссман. Жизнь и судьба, ч. 1 (1960)]

(4.14) Он *вернулся и топал* по палубе каблуками. [Б. С. Житков. «Мираж» (1935)]

(4.15) И бросился Христо в озеро, *лёг и грёб* под себя золото. [Б. С. Житков. Элчан-Кайя (1926)]

(4.16) Она постояла немного и грациозно, как балерина, прыгнула в другую сторону, опять *стала и глядела* всё в одну точку. [М. М. Пришвин. Кэт (1925)]

(4.17) Кондуктор *замолчал и оглядывал* его с ног до головы. [В. В. Вересаев. В степи (1901)]

(4.18) Прочитал и две минуты не мог ничего сказать, только *побагровел и сопел*, потом говорит: «Дайте карандаш!» — и тут же начертал резолюцию на телеграмме: «Чтоб духу его в Петербурге не было. [М. А. Булгаков. Записки покойника (Театральный роман) (1936–1937)] <возможно также понимание в значении одновременности>

(4.19) Силослав очень тому *удивился и спрашивал* их, откуда они. [М. Д. Чулков. Пересмешник, или Славенские сказки (1766–1768)]

(4.20) *Она *сняла и протирала* очки. [Александр Солженицын. В круге первом, т. 1, гл. 26–51 (1968) // «Новый Мир», 1990]
Надо сказать: *Она сняла очки и стала их протирать*.

В примерах (2.3а), (2.4а) из чешского языка глаголы НСВ как раз обозначают активную деятельность. Примеры (4.13)–(4.20) неудачны по той же причине: в некоторых контекстах глаголы НСВ активной деятельности не способны выступать в секвентной функции.

Для примера (4.21) предпочтительна интерпретация, при которой глагол НСВ имеет сдвинутый момент наблюдения, расположенный «в середине» имперфективной ситуации, но не в перфектном состоянии глагола СВ:

(4.21) Петр *выскочил и шел* по пояс в воде, своими руками помогая тащить судно. [А. С. Пушкин. История Петра: Подготовительные тексты (1835–1836)]

То же в примере (4.22):

(4.22) Опустившись на четвереньки, дети сидели в зале одни, но уже неумолимо приближалась команда уборщиц с зычными голосами, они *вошли и двигались* по проходам к сцене. Людмила Петрушевская. Маленькая волшебница // «Октябрь», 1996]

В отличие от примеров (4.13)–(4.20), здесь нет необходимости в выражении начинательности.

Однако однозначных свидетельств в пользу того, что глагол НСВ имеет в (4.21), (4.22) свой момент наблюдения, не локализованный в перфектном состоянии глагола СВ, нет. Формы *шел* из (4.21) и *двигались* из (4.22) выражают активную деятельность; тем не менее, начинательная фаза опущена, а ощущения шероховатости, как в примерах (4.13)–(4.20), не возникает.

5. Побочные факторы

Теперь общая картина обрисована, и можно обсудить детали.

5.1. Единое обстоятельство времени

Единый для СВ и НСВ момент наблюдения может быть задан обстоятельством времени. Тогда нет никакого сомнения в том, что момент наблюдения для имперфективной ситуации (стативной) приходится на середину перфектного состояния перфективной:

(5.1.1) Он теперь *разделся и стоял* в трусах. [Фазиль Искандер. Рассказ о море (1962)]

(5.1.2) Теперь этот офицер *пришёл и ждал* отца в кабинете. [Ю. О. Домбровский. Обезьяна приходит за своим черепом, часть 1 (1943–1958)] ≠ *стал ждать*.

(5.1.3) Но потом, когда *отужинали и сидели* сумерничая, она сказала: [Борис Екимов. Пиночет (1999)]

См. также примеры (2.1), (2.4) и пример (4.7).

(2.1) К тому моменту мы уже *подружились и питали* друг к другу нежные чувства. [Сати Спивакова. Не всё (2002)]

Так, в (2.1) одновременность событию; т. е. нет последовательности: сначала подружились, потом питали чувства. И это из-за общего обстоятельства.

Между тем в (5.1.4), (5.1.5) *теперь* стоит при глаголе НСВ и задает новый момент наблюдения для имперфективной ситуации. Здесь *теперь*, напротив, задает точку отсчета, сдвинутую с перфектного состояния глагола СВ и перенесенную вперед:

(5.1.4) Солнце постепенно *переместилось и било* теперь в правые окошки вместо левых. [И. Грекова. На испытаниях (1967)]

(5.1.5) Он сразу *преобразился и смотрел* теперь на нас несколько сверху, любовно, отечески строго. [Василий Аксенов. Пора, мой друг, пора (1963)]

Кажется, что в (5.1.6) есть подразумеваемое *теперь* — возникает новый момент наблюдения, не тот, который находится в перфектном состоянии глагола *переправили*:

(5.1.6) Деньги за квартиру они Насте *переправили и* <теперь> *собирались* уезжать. [Андрей Волос. Недвижимость (2000) // «Новый Мир», 2001]

В (5.1.7) у глагола НСВ определенно свой момент наблюдения, поскольку он имеет свое обстоятельство времени:

(5.1.7) хутор Зоричев, где *родились и жили* когда-то в ранней младости. [Борис Екимов. Пиночет (1999)]

5.2. Обстоятельство длительности

Обстоятельство длительности превращает длящийся процесс в законченный, т. е. в событие. Это объясняет безупречность сочетания СВ и НСВ в примерах (5.2.1)–(5.2.6) — возникает значение временной последовательности:

(5.2.1) Он тяжело *вздохнул и молчал*, наверное, целую минуту. [Андрей Геласимов. Ты можешь (2001)]

(5.2.2) Он снова *замолчал и молчал* так долго, что Яков спросил: [Ю. О. Домбровский. Факультет ненужных вещей, часть 4 (1978)]

(5.2.3) В этом счастливом состоянии я *уехал и пребывал* в нём до самого приезда Сони. [Анатолий Рыбаков. Тяжелый песок (1975–1977)]

(5.2.4) Кямал заснул и плакал во сне. [Токарева Виктория. Своя правда // «Новый Мир», 2002]

(5.2.5) Простить тех выношенных, намеренно-обидных слов невестке, которую, когда-то красивую, двадцатичетырехлетнюю, впервые пришедшую в их дом ещё старшеклассницей, мама так *полюбила и хранила* эту любовь очень долго, видеть её уязвлённая женщина больше не смогла. [Алексей Варламов. Купавна // «Новый Мир», 2000]

(5.2.6) Кошка страшно *обиделась и отказывалась* общаться недели две. [Иван Давыдов. Слова и кошки // «Русская жизнь», 2012]

5.3. Специальные глаголы

В (5.3.1)–(5.3.4) глаголы НСВ особые — они семантически близки к парному СВ (*продолжал* » *продолжил*, *просили* » *попросили*, *начинало* » *начало*), так что сочетание СВ+НСВ, со значением временной последовательности, так же законно, как сочетание двух СВ:

(5.3.1) Павел взглянул на неё, <…> секунду *помолчал и продолжал* говорить: [Андрей Волос. Недвижимость (2000) // «Новый Мир», 2001]

(5.3.2) Вилка самодовольно улыбнулась, вспоминая этот эпизод, но тут же снова *нахмурилась и продолжала*: [Светлана Коровина. Ее острейшее величество // «Трамвай», 1990]

(5.3.3) В проходной шестого подъезда его *остановили и просили* оставить портфель. [Людмила Улицкая. Казус Кукоцкого [Путешествие в седьмую сторону света] // «Новый Мир», 2000]

(5.3.4) Всё *поглотилось и начинало* уже усваиваться, кофе мелкими глотками довершал поздневечернюю трапезу вдовца, на экране заглушённого телевизора двигались и жестикулировали представители рабочего класса и научно-технической интеллигенции. [Анатолий Азольский. Лопушок // «Новый Мир», 1998]

В текстах Пушкина встречается несколько глаголов, употребленных в НСВ там, где сейчас нужна была бы форма СВ:

(5.3.5) Сергеев *прибыл и требовал* от нее выдачи изменника. [А. С. Пушкин. История Петра: Подготовительные тексты (1835–1836)]

(5.3.6) Батюшка к нему *подошел и благодарил* его с видом спокойным, хотя и тронутым. [А. С. Пушкин. Капитанская дочка (1836)]

(5.3.7) В сенях я *остановился и просил* у ней позволения ее поцеловать; Дуня согласилась… [А. С. Пушкин. Повести покойного Ивана Петровича Белкина/Станционный смотритель (5) (1830)]

(5.3.8) Шабашкин, видя, что он не в духе, *поклонился и спешил* удалиться. [А. С. Пушкин. Дубровский (1833)]

(5.3.9) Ибрагим *согласился и спешил* обратить разговор к предмету, более для него занимательному. [А. С. Пушкин. Арап Петра Великого (1828)]

(5.3.10) Вскоре я *выздоровел, и мог* перебраться на мою квартиру. [А. С. Пушкин. Капитанская дочка (1836)]

Тут налицо изменение нормы относительно выбора члена видовой пары.

5.4. У СВ и НСВ разные субъекты

В (5.4.1)–(5.4.4) у СВ и НСВ разные субъекты — обычно в этом случае союз И не выражает временной последовательности, так что момент наблюдения единый (это значение простого перечисления по Урысон 2011):

(5.4.1) К тому же дорогу *замело, и сыпал* снег. [Андрей Колесников. Бублики Мондео (2002) // «Автопилот», 2002.01.15]

(5.4.2) Здесь снег опять *кончился и зеленела* трава. [Фазиль Искандер. Святое озеро (1969)]

Однако в (5.4.3), (5.4.4) у глагола НСВ свой момент наблюдения. В (5.4.4) новый момент наблюдения задает *теперь*.

(5.4.3) мяч, попав в штангу, *отскочил и добивал* его в сетку кто-то другой. [И. Э. Кио. Иллюзии без иллюзий (1995–1999)]

(5.4.4) Деньги, заработанные на гастролях, быстро *кончились, и жили* они теперь с Лешей одной бедняцкой семьёй: [Людмила Улицкая. Казус Кукоцкого (Путешествие в седьмую сторону света) // «Новый Мир», 2000]

В (5.4.5) имперфективная ситуация синхронна перфектному состоянию перфективной:

(5.4.5) Печка давно *умолкла, и становилось* прохладно. [Виктор Ремизов. Воля вольная // «Новый мир», 2013]

5.5. Отрицание при глаголе СВ

Пропозиция с отрицанием часто не имеет временной локализации, поэтому неясно, тут единый для двух ситуаций момент наблюдения (на перфектном состоянии) или два, неизвестно как расположенных во времени. Значение союза здесь можно идентифицировать как И перечисления.

(5.5.1) Ведь в противном случае она государство *не обманула и получала* пенсии законно. [Сергей Николаев. Отцовство после смерти (2003) //«Богатей» (Саратов), 2003.11.20]

(5.5.2) Но Краснянский словно *не услышал и продолжал* горячо говорить про многомиллионные закупки зерна за рубежом как последствия коллективизации. [Василь Быков. Бедные люди (1998)]

(5.5.3) А я этого, дурак, *не понял и трепался*. [Ю. О. Домбровский. Факультет ненужных вещей, часть 2 (1978)]

(5.5.4) Отец, впрочем, *не обиделся и смеялся*. [Эдуард Лимонов. У нас была Великая Эпоха (1987)]

5.6. Имперфективная ситуация — продолжение перфектного состояния перфективной

Конфигурация «глагол СВ + союз И + глагол НСВ» часто реализуется в контексте, где глагол НСВ выражает ситуацию (состояние, процесс или деятельность), которая является **непосредственным продолжением** той ситуации, которая возникает в результате события СВ. (Примеры этого рода были и раньше.)

(5.6.1) Он *сел и сидел*, смотря на них всех, затаившийся, радостно-злой, готовый взорваться по первому поводу. [Ю. О. Домбровский. Факультет ненужных вещей, часть 1 (1978)]

(5.6.2) И она встала. Она *встала и стояла* на одной ноге, обняв одной рукой его за шею и пошатываясь. [Ю. О. Домбровский. Факультет ненужных вещей, часть 2 (1978)]

(5.6.3) Более того, на соседней верхней полке уже *улёгся и спал*, храпя и свесив волосатую руку с часами, какой-то мужик. [И. Грекова. В вагоне (1983)]

(5.6.4) Я сам *родился и рос* в Париже, [Иоанн Мейендорф. Православное свидетельство в современном мире (1992)]

(5.6.5) Он был бодрым и бдительным стражником, охранявшим жизнь непостижимого существа языка Платонова, которое *родилось и обитало* в нём. [Владислав Отрошенко. Эссе из книги «Тайная история творений» // «Октябрь», 2001]

(5.6.6) Володе в своё время поступали угрозы от Мовлади Удугова за то, что он *создал и поддерживал* антитеррористический сайт «Чечня. ру», — говорит Владимир Александрович. [Елена Лория. Владимира Сухомлина могли убить офицеры милиции (2003) // «Известия», 2003.01.12]

(5.6.7) Банк *купил и держал* ротмистр Кремнев. [Б. А. Пильняк. Простые рассказы (1923)]

В принципе, можно и здесь усмотреть синхронность имперфективной ситуации перфектному состоянию перфективной и, как производное, отношение временной последовательности.

6. Заключение

Исследование показало, что в пределах морфосинтаксической конфигурации «глагол СВ + союз И + глагол НСВ» временное соотношение достаточно хорошо предсказывается на базе внутрифразовой семантики вида — аналогично тому, как оно предсказывается в конфигурациях «глагол СВ + союз И + глагол СВ» и «глагол НСВ + союз И + глагол НСВ», см. Маслов 1984, Падучева 1996: 362–364.

Правило 1. Если у глагола НСВ момент наблюдения приходится на перфектное состояние глагола СВ, то возникает отношение **временной последовательности** между перфективной и имперфективной ситуацией, например:

(2.3) Такое сравнение он *придумал и гордился* его художественностью. [И. Грекова. Фазан (1984)]

(5.4.3) мяч, попав в штангу, *отскочил, и добивал* его в сетку кто-то другой. [И. Э. Кио. Иллюзии без иллюзий (1995–1999)]

Правило 2. Если у глагола НСВ момент наблюдения приходится на событие, обозначаемое глаголом СВ, то возникает отношение **одновременности** между перфективной и имперфективной ситуацией, например:

(2.1) К тому моменту мы уже *подружились и питали* друг к другу нежные чувства. [Сати Спивакова. Не всё (2002)]

Кроме того, значение одновременности возникает в ряде контекстов стативного глагола СВ:

(3.6) Рукав рубахи *оторвался и висел* на одной ниточке. [Ю. О. Домбровский. Обезьяна приходит за своим черепом, часть 3 (1943–1958)]

Имперфективная ситуация может не укладываться в перфектное состояние перфективной. Это возможно, в основном, в том случае, если глагол НСВ выражает деятельность. Здесь одно из двух.

а) Возникает аномалия или, во всяком случае, шероховатость, для устранения которой надо заменить глагол НСВ на соответствующий начинательный (или на сочетание с начинательным значением), см. примеры (4.13)–(4.20).

б) У глагола НСВ возникает свой момент наблюдения в середине имперфективной ситуации — начинательная фаза опущена, см. примеры (4.21), (4.22).

Впрочем, примеры (4.21), (4.22) могут трактоваться как подпадающие под Правило 1. Неясно, сдвинулся ли момент наблюдения при переходе от *выскочил* к *шел*, или остался прежним.

Не удается понять, почему в одних контекстах пропуск начальной фазы деятельности допустим, как в примерах (4.21), (4.22), а в других, как в примерах (4.13)–(4.20) для сдвига момента наблюдения желательно употребить глагол СВ, т. е. выразить начинательность.

Найти момент наблюдения, единый для перфективной и имперфективной ситуации или свой для каждой, — значит упорядочить их во времени. Однако временная упорядоченность может быть нерелевантна. Это имеет место в случае, если между ситуациями ощутима причинно-следственная связь, как в примерах (2.2) и (4.10)–(4.12).

Подтверждается тезис о том, что глагол НСВ в актуально-длительном значении в русском языке предполагает момент наблюдения, расположенный «в середине» ситуации: примеры (4.13)–(4.20) демонстрируют невозможность употребления НСВ в ингрессивном значении в русском языке. Процитирую фрагмент из Падучева 1996: 364. «Сопротивление ингрессивной интерпретации отличает русский НСВ прош. от многих претеритальных форм других языков, например, от простого прошедшего в английском; ср. пример из Carlson

1981: *At sunrise I walked eastward* — букв. ‘Когда солнце встало, я шел на восток’; по-русски надо сказать *я пошел на восток*.»

Стоит сравнить соотношения, выявленные в конфигурации «глагол СВ + союз И + глагол НСВ», со значениями союза И, предъявленными в Урысон 2011.

Отношение временной последовательности — это И нормального развития повествования по Урысон 2011, например:

(2.3) Такое сравнение он *придумал и гордился* его художественностью. [И. Грекова. Фазан (1984)]

Причинно-следственное отношение — это И нормального следствия по Урысон 2011, примеры (4.10)–(4.12):

(4.12) Станцию почти каждый день обстреливали тяжёлые немецкие орудия, — немцы наловчились, *пристрелялись и лепили* снаряды метко, по стенам цехов, грохот разрывов то и дело потрясал землю. [Василий Гроссман. Жизнь и судьба, ч. 1 (1960)]

Одновременность — это простое перечисление, чистая конъюнкция по Урысон 2011: 282, 292. Примеры:

(4.9) Ох ты господи, да Метальников, к вашему сведению, ещё до революции во Францию *уехал и заведовал* в Институте Пастера отделом. [Даниил Гранин. Зубр (1987)]

(5.4.2) Здесь снег опять *кончился и зеленела* трава. [Фазиль Искандер. Святое озеро (1969)]

Итак, показано, что разные временные соотношения возникают на базе видовых значений СВ и НСВ во внутрифразовом контексте и существенно опираются на понятие момент наблюдения. В выражении этих отношений принимают участие также временные адвербиалы. Можно думать, значение союза И во всех контекстах одно и то же.

Работа проводилась на базе НКРЯ. Но использованы далеко не все возможности Корпуса — глагол СВ и глагол НСВ могут быть отделены друг от друга подчиненными словами, которые находятся справа и слева от союза. Выборка оказалась достаточно представительной, чтобы понять общие принципы. Однако более широкий контекст может добавить новых подробностей.

Литература

1. *Glovinskaja M. Ja.* (1982), Semantic types of aspectual oppositions in the Russian verb [Semantičeskie tipy vidovyx protivopostavlenij russkogo glagola], Nauka, Moscow.

2. *Knjazev Ju. P.* (2014) Ingressive uses of Imperfective in Russian [Ingressivnye upotreblenija nesoveršennogo vida v russkom jazyke]. The Heritage of Juri Maslov: Linguistic Ideas and their Evolution [Nauchnoe nasledie i razvitie idej Ju. S. Maslova], St. Petersburg, 96–99.
3. *Paducheva E. V.* (1996) Semantic investigations. Semantics of tense and aspect in Russian. Semantics of narrative. [Semantičeskie issledovanija. Semantika vremeni i vida v Russkon jazyke. Semantika narrativa] Jazyki russkoj kul'tury, Moscow. http://lexicograph.ruslang.ru/TextPdf1/PaduSemantIssl1996.pdf
4. *Paducheva E. V.* (2010) Mirror symmetry of past and future: the figure of observer [Zerkal'naja simmetrija prošedšego i buduščego: figura nabljudatelja], Izvestija RAS. Literature and language series, v. 69 № 3, 16–20. http://lexicograph.ruslang.ru/TextPdf2/symmetr-2010.pdf
5. *Petruxina E. V.* (2000) Aspectual verb categories in Russian in comparison with Check, Slovak, Polish and Bulgarian [Aspektual'nye kategorii v russkom jazyke v sopostavlenii s češskim, slovackim, pol'skim i bolgarskim jazykami], MGU, Moscow.
6. *Uryson E. V.* (2011) An attempt at describing semantics of conjunctions [Opyt opisanija semantiki sojuzov]. Jazyki russkoj kul'tury, Moscow.
7. *Bäuerle R.* (1979) Temporale Deixis, temporale Frage. Tubingen: Gunter Narr, 1979.
8. *Carlson G.* (1981) Aspect and quantification. In: Syntax and Semantics, vol. 14. Tense and Aspect. N.Y. etc.: Acad. press, , 31–64.
9. *Kratzer A.* (1978) Semantik der Rede. Kronberg: Skriptor.
10. *Smith C. S.* (1997) Smith The parameter of aspect. 2d ed. Dordrecht 1997.

БАЗА ДАННЫХ ДЛЯ ИССЛЕДОВАНИЯ ВАРИАТИВНОСТИ ТВЕРДЫХ/МЯГКИХ СОГЛАСНЫХ ПЕРЕД *Е* В ЗАИМСТВОВАННЫХ СЛОВАХ[1]

Перова Д. М. (dmperova@mail.ru),
Бондаренко К. Е. (moidom@mail.ru),
Добрушина Н. Р. (nina.dobrushina@gmail.com)

НИУ ВШЭ, Москва, Россия

Работа посвящена вариативности твердых/мягких согласных перед *е* в заимствованных словах (*ка[ф]е*). Целью данного проекта является создание обширной базы слов, в которых встречается эта вариативность. База включает в себя словник, созданный с помощью анализа запросов в поисковой системе Яндекс. Все единицы словника имеют разметку по ряду параметров, которые необходимо учитывать при экспериментальном исследовании тенденций варьирования слов. В статье описан процесс формирования списка слов с чередованием твердого/мягкого согласного и принципы разметки списка. В настоящий момент слова размечаются по чередующемуся согласному, позиции сочетания в слове (первый слог, середина слова, конец слова), позиции по отношению к ударению, открытому/закрытому слогу, времени появления слова в русском языке (самый ранний случай употребления в Национальном корпусе русского языка); языку, из которого было заимствовано слово; частотности употребления слова (в выдаче поисковых систем Яндекс и Google, в списке запросов Яндекс, в НКРЯ). База будет использована при отборе слов-стимулов для проведения экспериментов, нацеленных на установление частотности вариантов в современной устной речи и выявление социальных коррелятов произнесения того или иного варианта (возраст, пол, образование и другие характеристики говорящих).

Ключевые слова: твердость/мягкость согласных, социолингвистика, заимствованные слова, социофонетика, вариативность

[1] Исследование выполнено при поддержке гранта РНФ № 16-18-02071.

A DATABASE FOR VARIATIONAL STUDIES OF PALATALIZED/VELARIZED CONSONANTS PRECEDING [E] IN LOANWORDS

Perova D. M. (dmperova@mail.ru),
Bondarenko K. E. (moidom@mail.ru),
Dobrushina N. R. (nina.dobrushina@gmail.com)

HSE, Moscow, Russia

The paper presents the initial/preparatory stage of the study of variation of hard/soft consonants before e in loanwords (*ka[f]e*). The main goal is to compile a database of relevant words for use in sociolinguistic research. The database is based on the list of word forms containing relevant contexts in users' queries to Yandex. All entries in the database are annotated for parameters that may be important in a variational study of the phenomenon. The article describes how the list was compiled and the principles of its annotation. The latter includes the consonant, the position of the consonant re the stressed syllable, the type of syllable where it occurs (open/closed), the year of the first occurrence of the word in Russian National Corpus; the language from which it was borrowed; its frequency. The database may be used to select stimuli for experimental studies of variation in modern speech and of its social correlates (age, gender, education, etc).

Keywords: hard/soft consonants, sociolinguistics, loanwords, sociophonetics, variation

1. Вариативность твердых/мягких согласных перед *е*

В заимствованных словах наблюдается вариативность твердого/мягкого согласного перед *е*. Так, общепринятым является твердое произнесение *т* в слове *компьютер* и *н* в слове *сонет*; при этом мягкое произнесение этих согласных тоже встречается нередко. Твердый согласный представляет собой фонетический «след» того языка, где противопоставление по твердости/мягкости отсутствует. Таким образом слово сохраняет произношение, соответствующее произношению в языке-источнике.

Известно, что твердое произнесение согласного в этой позиции появилось уже в XIX веке. М. В. Панов пишет, что для второй половины XIX века характерно «тактично, „не переигрывая", сохранять признаки заимствованности в определенном кругу слов и избегать их за пределами этого круга» (Панов 1990: 148).

Поначалу, по-видимому, твердое произношение было характерно для образованных кругов общества («...первоначально оно появилось в речи интеллигенции, чиновничества, высших классов, знающих иностранные

языки» — Аванесов 1984). Вариативность, тем самым, сразу носила ярко выраженный социальный характер.

Высказывались предположения о том, что тенденция к твердому произношению сойдет на нет. Во-первых, такой представлялась общая логика освоения заимствований (со временем адаптироваться к фонетике языка-реципиента), а во-вторых, социальные катаклизмы первой трети XX века значительно снизили роль интеллигентной части общества, в речи которой наблюдалась тенденция к сохранению твердого произнесения в соответствии с фонетикой языка-источника, и вывели на первый план речь того социального слоя, которому иностранные языки были незнакомы. Однако эти прогнозы не подтвердились. В последней трети XX века стало ясно, что произношение с твердым согласным не только не ушло, но и стало охватывать все более широкий круг позиций. Современные исследования показывают, что сочетания, которые были недопустимыми еще недавно, сегодня распространены в речи молодежи (Аванесов 1984, Касаткин 2000, Каленчук, Касаткина 2011: 108–111, Стаферова 2014).

Хотя вариативность твердых/мягких согласных неоднократно становилась предметом исследования (Гловинская 1971, Князев 1987, Краснова, Смирнова 2012), детального представления о том, какова динамика развития этого явления, у нас до сих пор нет. Есть слова, для которых можно отметить тенденцию к освоению, то есть к повышению частотности мягкого произнесения. В работе С. В. Князева сказано, что в слове *миксер* произносится твердый согласный (Князев 1987), а в словаре (Каленчук, Касаткин, Касаткина 2012) это слово приводится уже с мягким. С другой стороны, есть слова, которые сохраняют тенденцию к твердому произношению на протяжении всей своей наблюдаемой жизни (*темп, энергия*). Есть и такие, для которых твердое произношение стало более распространенным, чем раньше (*артерия, бактерия* — Князев 1987). Кроме того, мощный поток новых заимствований и усилившаяся в самое последнее время роль английского языка в русской речевой культуре модифицируют условия существования заимствованных слов.

Почему некоторые слова адаптируются к смягчению согласных перед *е* в русском языке, а другие сохраняют твердость или даже усиливают тенденцию к произнесению твердого согласного? Какими факторами определяется различное поведение слов с вариативностью твердого/мягкого согласного перед *е*? На этот счет существует ряд гипотез; они будут приведены в разделе 4. Для того чтобы подтвердить или опровергнуть эти гипотезы, необходимо систематическое исследование вариативности произношения заимствованных слов. Такое исследование должно опираться на количественные данные о том, какой процент говорящих с определенными социальными характеристиками (прежде всего имеет значение возраст) произносит то или иное слово с твердым согласным. Данные для исследования могут быть получены лишь в результате эксперимента, поскольку устных корпусов достаточного объема в настоящее время не существует.

В этой работе будет представлен проект создания базы слов, содержащих чередование твердого/мягкого согласного перед *е* с разметкой; база послужит источником материала для проведения экспериментов.

2. Цели и задачи проекта

Целью проекта является создание базы заимствованных слов с вариативным произношением согласных перед *е,* размеченных по ряду признаков и параметров.

На сегодняшний день списки таких слов можно почерпнуть из двух источников.

Во-первых, это исследовательские работы, посвященные вариативности твердых/мягких согласных перед *е* (например, Князев 1987, Каленчук, Касаткина 2013, Ларионова 2014). Ни одна из этих работ, однако не содержит полного списка, в частности потому, что не охватывает современной лексики (такой, как *фреш, френдить, флешмоб*) и имен собственных (*Фредди Меркьюри, Ангела Меркель*).

Вторым источником являются орфоэпические словари, которые тоже содержат современную лексику в весьма ограниченном количестве. Например, в словаре (Каленчук, Касаткин, Касаткина 2012) нет слов *флешка, хипстер, фейк*, нет и имен собственных.

Основным назначением такого списка является отбор слов для экспериментальных исследований в области твердого/мягкого произношения согласного. Для того чтобы эксперимент подтвердил или опроверг гипотезу о влиянии определенного фактора на произношение согласного, необходимо отбирать слова так, чтобы они различались лишь одним признаком (например, только типом согласного, после которого находится *е*). Поэтому все слова в базе размечаются по ряду параметров. В основу разметки легли существующие в современной лингвистике гипотезы о том, что именно может влиять на поведение согласного (например, тип согласного, ударение и др. — см. раздел 4).

Таким образом, задачами этого исследования было

а) составить большой список, который в результате содержал бы 3000–4000 релевантных слов (раздел 3);

б) составить список факторов, от которых может зависеть вариативность в иностранных словах с *е* (раздел 4);

в) разметить слова по выбранным параметрам (раздел 5).

3. Составление предварительного списка слов

Для составления списка заимствованных слов с релевантными сочетаниями было решено использовать два источника. На начальном этапе составление словника велось с помощью списка запросов в Яндекс с вариативным написанием через *е* или через *э*. На следующем этапе работы планируется привлечь данные орфоэпического словаря русского языка (Каленчук, Касаткин, Касаткина 2012)[2].

В качестве основного источника был выбран список запросов в поисковой системе Яндекс. Предполагалось, что если в слове имеется вариативность

[2] Авторы выражают благодарность М. Л. Каленчук за предоставление электронной версии словаря.

твердого/мягкого согласного, то среди поисковых запросов будут присутствовать оба варианта. Так, если много людей произносят слово *термин* с твердым [т], то кто-то из них может ошибиться, написав в строке запроса *тэрмин* через «э».

Был получен список всех запросов, которые различаются лишь буквами *е* или *э*[3]. Список представлял собой около 280 000 поисковых запросов пользователей Яндекса в формате «Слово — частотность». Список, однако, содержал слишком много нерелевантных для нашей задачи запросов:

Табл. 1. Фрагмент из списка запросов в Яндекс и их частотности

сканвоквер	44
сканвоквэр	122
сканворде	23423
сканвордэ	69
сканвэй	1077
скане	502
сканекс	5329
сканер	3973556

Для того чтобы избавиться от нерелевантных строк, были проведены следующие операции[4]:

- применение фильтров для очистки списка от «мусора»: цифры, некириллические символы, слова с дефисом, «э» или «е» после букв *й*, *ц*, *у*, *е*, *ш*, *щ*, *ъ*, *ы*, *а*, *о*, *ж*, *э*, *я*, *ч*, *и*, *ь*, *ю*, *ъ*, некоторые опечатки,
- составление пар запросов по написанию «э» или «е» и исключение из списка непарных запросов,
- применение частотного фильтра: частотность каждого запроса в паре больше или равна 200; запросы с такой частотностью никогда не содержали релевантных для задачи проекта слов.

Итогом стал список (см. Табл. 2), состоящий приблизительно из ~28 000 строк в формате «word1 + freq1 + word2 + freq2», где freq1 и freq2 — это частотности соответствующих запросов.

В полученном списке всё равно оказалось значительно больше «мусора», чем реальных слов с вариативностью твердого/мягкого согласного перед *е*. Для того чтобы более точно оценить процент ненужных строк в этом списке, был проведен эксперимент с использованием метода blind annotation (слепая аннотация). Несколько аннотаторов[5] должны были выделить в случайно сформированных

---

[3] Выражаем благодарность компании Яндекс за предоставленные данные по запросам и в особенности сотрудникам Яндекса в лице Григория Носырева за их подготовку и помощь в обработке.

[4] В обработке списка принимал активное участие М. А. Даниэль.

[5] В аннотировании участвовали К. Бондаренко, М. Бородина, Н. Добрушина и Д. Перова.

из списка выборках в 500 слов все нерелевантные запросы, а в случайно сформированной выборке в 100 слов — 20 нужных. Из 500 слов оказалось 462 таких, которые были помечены как ненужные хотя бы одним аннотатором. Выбрать из 100 случайных слов 20 нужных запросов никому из аннотаторов не удалось, потому что их было значительно меньше. Таким образом стало очевидным, что список должен быть сокращен более чем в пять раз.

Табл. 2. Фрагмент из предварительного списка в 28 000 строк

сканер	3973556	сканэр	1523
скарфейс	9091	скарфэйс	740
скатмен	1200	скатмэн	346
скачатькеш	368	скачатькэш	289
скачатьреп	1556	скачатьрэп	914
скачатьторент	30899	скачатьторэнт	341
скачатьфлеш	2994	скачатьфлэш	364
скачатье	3863	скачатьэ	30865
скачате	16684	скачатэ	1321
сквирленд	660	сквирлэнд	203
сквирел	73615	сквирэл	828
скве	8482	сквэ	623
сквеа	4628	сквэа	1101

Для того чтобы в список попали даже те слова, которые не включены в словари, было принято решение проводить дальнейшую чистку вручную. В настоящий момент список насчитывает около ~1500 имен нарицательных и собственных, которые делятся на следующие категории:

- имена нарицательные: ~800;
- имена собственные (имена и фамилии людей, топонимы и т. д.): ~500;
- имена собственные (названия брендов, компаний, продуктов и т. д.): ~200.

Данные из Яндекса дают ценную информацию о частотности различных запросов. В частности, видно, что для большинства слов более частотным является запрос с буквой *е*:

Табл. 3. Слова из списка Яндекса с высокой частотностью написания через *е*

Запрос с *е*	**Частотность**	**Запрос с *э***	**Частотность**
адаптер	5 586 492	адаптэр	1 268
ассемблер	162 704	ассэмблер	204
баннер	1 426 581	баннэр	278
гангстер	424 047	гангстэр	250
конвертер	4 728 557	конвертэр	209
музей	17 975 495	музэй	875
ортез	323 939	ортэз	531

Есть и слова, у которых частотнее запрос с *э* (Табл. 4).

Табл. 4. Слова из списка Яндекса с высокой частотностью написания через *э*

Запрос с *е*	**Частотность**	**Запрос с *э***	**Частотность**
флешмоб	643 334	флэшмоб	766 621
фидбек	16 246	фидбэк	27 374
сендвич	453 163	сэндвич	2 050 457
семплинг	12 671	сэмплинг	20 499
слеш	226 202	слэш	757 050
меппинг	3 743	мэппинг	11 367
минивен	299 520	минивэн	541 607

Далеко не всегда частотность варианта с *е* или с *э* отражает частотность соответствующего произношения. Написание находится под сильным влиянием устоявшейся орфографии, а на те слова, которые реже встречаются в сфере нормализованной письменной речи, правила будут влиять в меньшей степени (*слеш, минивен*). Имеет значение и часть слова: элемент *ер* в словах *адаптер, баннер, гангстер*, по-видимому, пишется по аналогии со многими другими словами, имеющими этот элемент. Таким образом, в большом количестве случаев значимым для исследования является не столько отношение частотности варианта с *э* к частотности запроса с *е*, сколько абсолютная частота запроса с *э*.

Следующий этап работы над пополнением списка будет вестись с использованием данных орфоэпических словарей, в первую очередь — словаря Каленчук, Касаткин, Касаткина 2012. Нас будут интересовать все словарные статьи, которые содержат любую оценку варианта произношения согласного перед *е*, в частности как неправильного (пример (1)), допустимого (пример (2)), равноправного (пример (3)):

(1) *АВТОПОРТРЕТ, автопортре́та, мн. автопортре́ты, автопортре́там* \\ *автопо[ртр'е́]т* ! ***неправ****. автопо[ртрэ́]т*

(2) *АВТОЦИСТЕРНА, автоцисте́рны, мн. автоцисте́рны, автоцисте́рнам,может произноситься с дополнительным ударением: а̀втоцисте́рна* *автоци[с'т'е́]рна и* ***допуст.*** *автоци[стэ́]рна*

(3) *АБРЕК, абре́ка, мн. абре́ки, абре́кам* \\ *а[бр'е́]к* ***и*** *а[брэ́]к*

В дальнейшем будут привлечены и другие словари, в той или иной степени отражающие произношение заимствованных слов.

4. Факторы, влияющие на произнесение согласного

Список слов с вариативностью согласного по твердости/мягкости размечается в соответствии с набором факторов, релевантных для произношения согласного. В предшествующих исследованиях был отмечен ряд явлений, которые могут влиять на произношение согласного.

Наиболее значимым, по-видимому, является тип согласного. Во всех работах на эту тему указывалось, что сохранение твердости согласного характерно в наибольшей степени для зубных согласных (*т, д, с, з, н, р*) и в наименьшей — для задненебных согласных (*г, к, х*) (Панов 1990, Аванесов 1984). Все слова в базе размечены по согласному.

Известно также, что играет роль место слога по отношению к ударению. Так, различаются произношения слов *се́кс* (обычно твердый согласный) и *сексуа́льный* (чаще встречается мягкое произношение) (Князев, Пожарицкая 2005: 239). Кроме того, для некоторых слов в базе характерна вариативность ударения: *фе́йсбук* или *фейсбу́к*. В этом случае отмечаются оба варианта.

Особенно располагает к сохранению твердости согласного позиция в открытом ударном слоге в конце слова (*пюре́, кафе́*). Эта тенденция отмечена как одна из наиболее ярких во многих работах (Князев, Пожарицкая 2005, Ларионова 2014). В базе помечаются открытые/закрытые слоги.

Наконец, по-видимому, особую роль играет суффиксоидный показатель *-ер*. Есть свидетельства того, что слова, имеющие этот элемент, дольше сохраняют твердость: *свитер, компьютер, лазер* (Стаферова 2014). Слова, имеющие этот элемент, мы помечаем отдельно.

Возможно, играет роль время, когда слово появилось в русском языке. Ка жется естественным, что более старые заимствования должны иметь более выраженную тенденцию к смягчению, в то время как более новые чаще сохраняют твердость. Это предположение, однако, требует систематической экспериментальной проверки. Реальная ситуация, по-видимому, сложнее. Как указывает Е. Ю. Ларионова, слово *конгресс* однозначно воспринимается говорящими как иностранное, в то время как слово *свитер* многие считают исконно русским, несмотря на то, что первое вхождение слова *конгресс* в НКРЯ датируется 1768 годом, а слова *свитер* — 1923 (Ларионова 2014).

Таким образом, «возраст» слова нужно учитывать наряду с другими параметрами. Одним из них является частотность: есть исследования, которые показывают, что более частотные слова осваиваются быстрее, чем редкие (Poplack & Sankoff 1984). В экспериментальном исследовании произношения сочетания *ме* в именах собственных, проведенном группой студентов, обнаружилась корреляция между частотностью мягкого произнесения *м* и частотностью упоминания слова в поисковых системах (в слове *Ангела **Ме**ркель* частотность мягкого произнесения «м» оказалась наиболее высокой) (Антонова, Кустова, Пекарская, Полянская 2015). Таким образом, частотность должна учитываться при составлении списка слов для экспериментов: слова с разной частотностью могут дать разный результат при одинаковом согласном, месте ударения и других свойствах. В базе отмечается частотность слова в НКРЯ, в поисковых системах Яндекс и Google, причем для Яндекса имеются как данные о частотности запросов, так и данные о частотности упоминания.

- В литературе встречаются также наблюдения о том, что успешность адаптации заимствований связана с тем, из какого языка пришло слово; язык-источник, однако, имеет значение не сам по себе, а в связи с его

престижностью в определенной сфере (Lev-Ari & al. 2014). Поэтому слова в базе планируется разметить по языку-источнику, хотя не во всех случаях такая информация доступна и достаточно надежна.

5. Разметка слов в базе

Итак, база представляет собой таблицу, в которой каждому слову по каждому параметру присвоено некоторое значение (см. Табл. 5). В настоящий момент производится разметка по следующим параметрам:

- согласный;
- позиция сочетания в слове: первый слог, середина слова, конец слова.
- ударение в слове: тип слога, в котором находится интересующее нас сочетание «согласный + Е» — ударный, заударный, предударный и т. п.; отмечается также вариативность ударения;
- открытый/закрытый слог;
- время появления слова в русском языке (самый ранний случай употребления в Национальном корпусе русского языка);
- язык, из которого было заимствовано слово;
- частотность употребления слова (в выдаче поисковых систем Яндекс и Google, в списке запросов Яндекс, в НКРЯ).

Табл. 5. Отрывок таблицы с параметрами заимствованных слов

Word	Согласный	Заимствование	Ударение	Вариативность ударения	Слог	Позиция в слове	слова с - ер	Яндекс-частотность				НКРЯ-частотно	Первое вхожден
								е, (е)	э, (е)	е, э	э, э		
адаптер	т	англ	заударный слог	невар	закрытый	конечный слог	ер	5586492	1268			105	1926
ассемблер	с	англ	ударный слог	невар	закрытый	середина слова	ер	162704	204			9	1991
аутлендер	л	англ	ударный слог	невар	закрытый	середина слова	ер	2742181	18098	673	3121	14	2004
аутлендер	д	англ	заударный слог	невар	закрытый	конечный слог	ер	2742181	18098	673	3121	14	2004
баннер	н	англ	заударный слог	невар	закрытый	конечный слог	ер	1426581	278			152	2000
битмейкер	м	англ	ударный слог	невар	закрытый	середина слова	ер	19811	1013			0	
бластер	т	англ	заударный слог	невар	закрытый	конечный слог	ер	332875	1555			15	1998
блейзер	л	англ	ударный слог	невар	закрытый	начало слова	ер	497547	24321			55	1979
блендер	л	англ	ударный слог	невар	закрытый	начало слова	ер	2574051	16253	1026	2761	57	1999
блендер	д	англ	заударный слог	невар	закрытый	конечный слог	ер	2574051	16253	1026	2761	57	1999
браузер	з	англ	заударный слог	невар	закрытый	конечный слог	ер	15554038	2524			110	1997
бустер	т	англ	заударный слог	невар	закрытый	начало слова	ер	717491	460			14	1998
газгольдер	д	англ	заударный слог	невар	закрытый	конечный слог	ер	1980269	500			14	1939
гангстер	т	англ	заударный слог	невар	закрытый	конечный слог	ер	424047	250			332	1997
дисклеймер	л	англ	ударный слог	невар	закрытый	середина слова	ер	28779	418			2	2011
имиджмейкер	м	англ	ударный слог	невар	закрытый	середина слова	ер	53662	222			148	1996
кластер	т	англ	заударный слог	невар	закрытый	конечный слог	ер	586431	359			816	1974
клинер	н	англ	заударный слог	невар	закрытый	конечный слог	ер	481808	352			0	
клипмейкер	м	англ	ударный слог	невар	закрытый	середина слова	ер	13577	200			49	1997
клифхенгер	к	англ	ударный слог	невар	закрытый	середина слова	ер	738	3812			0	
компьютер	т	англ	ударный слог	невар	закрытый	конечный слог	ер	70530841	2964			8520	1969
конвертер	т	англ	заударный слог	невар	закрытый	конечный слог	ер	4728557	209			98	1956

Наличие в базе слов разметки по названным выше параметрам позволит исследователю подбирать для эксперимента слова, различающиеся только по одному параметру. Например, если исследователя интересует зависимость произношения от согласного, достаточно выбрать одинаковые значения других параметров. Допустим,

- заимствование: английский язык,
- ударный слог,
- закрытый слог,
- конечный слог,
- не *-ер*,
- заимствование не раньше 1990-х годов.

Получим следующие слова: *ап**грейд**, ап**дейт**, гео**тег**, и**мейл**, интер**нет**, интер**фейс**, кос**плей**, микс**тейп**…* и т. д.

6. Конечный результат

Конечный продукт данного проекта будет представлять собой структурированную базу для исследований вариативности твердых/мягких согласных в позиции перед *е* в заимствованных словах русского языка — то есть наиболее полный список релевантных заимствованных слов, размеченных по лингвистическим параметрам.

База облегчит подбор материала для социолингвистических экспериментов. По мере осуществления проекта предполагается проверка корреляции произношения твердого/мягкого согласного и ряда экстралингвистических параметров, а именно возраст говорящего, гендер, уровень образования, знакомство с иностранными языками, род деятельности (может влиять на степень освоенности некоторых заимствованных слов в определенных группах информантов), восприятие слова как «иностранного» и другие.

Кроме того, данные поисковых запросов в Яндексе могут быть использованы для подготовки новых версий орфоэпических словарей русского языка.

Литература

1. *Аванесов Р. И.* (1984), Русское литературное произношение, 6 изд., Просвещение, Москва.
2. *Антонова Е., Кустова М., Печникова В., Полянская Л.* (2015), Смягчение [м] перед [э]/[е] в заимствованных именах собственных, Антропология. Фольклористика. Социолингвистика. Конференция студентов и аспирантов, Санкт-Петербург.
3. *Гловинская М. Я.* (1971), Об одной фонологической подсистеме в современном русском литературном языке, Развитие фонетики современного русского языка: Фонологические подсистемы, Наука, Москва.
4. *Каленчук М. Л., Касаткин Л. Л., Касаткина Р. Ф.* (2012), Большой орфоэпический словарь русского языка, АСТ-Пресс, Москва.
5. *Князев С. В.* (1987), Корреляция согласных по твёрдости/мягкости в современной орфоэпической норме, Русский язык как иностранный в отраслевом вузе, Наука.
6. *Князев С. В., Пожарицкая С. К.* (2005), Современный русский литературный язык. Фонетика. Графика. Орфография. Орфоэпия., Академический Проект, Москва.
7. *Краснова Е. В., Смирнова Н. С.* (2012), Региональные предпочтения в произношении некоторых русских и заимствованных слов, Компьютерная лингвистика и интеллектуальные технологии, Изд-во РГГУ, Москва, С. 307–318.

8. *Ларионова Е. Ю.* (2015), Вариативность произношения согласных в заимствованных словах как социальная переменная: диссертация на соискание степени магистра. Направление подготовки 035800 «Фундаментальная и прикладная лингвистика», Европейский университет в Санкт-Петербурге, Санкт-Петербург, 61 л.
9. *Панов М. В.* (1990), История русского литературного произношения XVIII–XX вв., Наука, Москва, С. 453.
10. *Поливанов Е. Д.* (1931), О фонетических признаках социально-групповых диалекты и в частности русского стандартного языка, За марксистское языкознание, Наука, Москва, С. 124–137.
11. *Стаферова Д. А.* (2014), Социолингвистическое исследование вариативности твёрдости согласного [Т] или [Т'] перед гласным Е, Русский язык в научном освещении: научный журнал, №. 2., С. 104–125.
12. *Lev-Ari, S., San Giacomo, M., & Peperkamp, S.* (2014). The effect of domain prestige and interlocutors' bilingualism on loanword adaptations. Journal of Sociolinguistics, 18(5), 658–684.
13. *Poplack, Shana and David Sankoff.* (1984) Borrowing: The synchrony of integration. Linguistics 22: 99–135.

References

1. *Avanesov R. I.* (1984), Russian standard pronunciation [Russkoe literaturnoe proiznoshenie], 6th ed., Prosveschenie, Moscow.
2. *Antonova E., Kustova M., Pechnikova V., Polyanskaya L.* (2015), Palatalization of [m] followed by [e]/[é] in borrowed proper names [Smyagchenie [m] pered [e]/[é] v zaimstvovannih imenah sobstvennih], Anthropology. Folklore. Sociolinguistics. Students and postgraduates' conference [Antropologia. Folkloristika. Sociolingvistika. Konferencia studentov i aspirantov], St. Petersburg.
3. *Glovinskaya M. Y.* (1971), A phonological subsystem standard Russian language: a case study [Ob odnoi fonologicheskoi podsisteme v sovremennom russkom literaturnom yazyke], Phonetic evolution of modern Russian: Phonological subsystems [Razvitie fonetiki sovremennogo russkogo yazika: Fonologicheskie podsistemy], Nauka, Moscow.
4. *Kalenchuk M. L., Kasatkin L. L., Kasatkina R. F.* (2012), A comprehensive dictionary of Russian orthoepics [Bol'shoi orfoepicheskii slovar' russkogo yazyka], AST-Press, Moscow.
5. *Knyazev S. V.* (1987), Velarized-palatalized opposition of consonants in standard Russian [Korrelyacia soglasnih po tvyordosti/myagkosti v sovremennoi orfoepicheskoi norme], Russian as a foreign language in specialized colleges [Russkii yazik kak inostrannyi v otraslevom vuze], Nauka.
6. *Knjazev S. V., Pozharickaja S. K.* (2005), Modern standard Russian. Phonetics. Graphics. [Sovremennyj russkij literaturnyj jazyk. Fonetika. Graphics. Orthography. Orthoepy], Akademicheskij Proekt, Moskva.

7. *Krasnova E. V., Smirnova N. S.* (2012), Regional tendencies in the pronounciation of some Russian loanwords [Regional'nye predpochtenija v proiznoshenii nekotoryh russkih i zaimstvovannyh slov], Komp'juternaja lingvistika i intellektual'nye tehnologii, Izd-vo RGGU, Moskva, p. 307–318.
8. *Larionova E. Ju.* (2015), Consonant realization in loanwords as a socially determined variable. MA dissertation. [Variativnost' proiznoshenija soglasnyh v zaimstvovannyh slovah kak social'naja peremennaja: dissertacija na soiskanie stepeni magistra.] Evropejskij universitet v Sankt-Peterburge, Sankt-Peterburg.
9. *Lev-Ari, S., San Giacomo, M., & Peperkamp, S.* (2014). The effect of domain prestige and interlocutors' bilingualism on loanword adaptations. Journal of Sociolinguistics, 18(5), 658–684.
10. *Panov M. V.* (1990), History of Russian standard pronunciation in XVIII–XX. [Istorija russkogo literaturnogo proiznoshenija XVIII–XX vv.], Nauka, Moskva.
11. *Polivanov E. D.* (1931), On phonetics of some social dialects in Russian standar language. [O foneticheskih priznakah social'no-gruppovyh dialekty i v chastnosti russkogo standartnogo jazyka], Za marksistskoe jazykoznanie, Nauka, Moskva, p. 124–137.
12. *Poplack, Shana and David Sankoff.* (1984) Borrowing: The synchrony of integration. Linguistics 22: 99–135.
13. *Staferova D. A.* (2014), A study in sociophonetic variation: palatalized vs. Non-palatalized [t] followed by [e]. [Sociolingvisticheskoe issledovanie variativnosti tvjordosti soglasnogo [T] ili [T'] pered glasnym E], Russkij jazyk v nauchnom osveshhenii: nauchnyj zhurnal, № 2, S. 104–125.

INTRA-SPEAKER STRESS VARIATION IN RUSSIAN: A CORPUS-DRIVEN STUDY OF RUSSIAN POETRY

Piperski A. Ch. (apiperski@gmail.com)

Russian State University for the Humanities / National Research University Higher School of Economics, Moscow, Russia

Kukhto A. V. (anton.kukhto@gmail.com)

Lomonosov Moscow State University, Moscow, Russia

Russian lexical stress exhibits both inter-speaker variation, defined by the speaker's regional affiliation, social status, age, etc., as well as intra-speaker variation. The latter is difficult to capture due to the need for large corpora of spoken text produced by one speaker. These are lacking, but can be replaced with poetic corpora. We use automatic analysis of poetic texts by 10 poets, drawn from the Russian National Corpus, in order to find word forms that can have stress variation. The number of such forms for an individual speaker ranges between 30 and 200 words, distributed among different parts of speech. We propose a quantitative measure of overall stress variability independent of the corpus size and show that there is a tendency for this variability to diminish over time, at least in poetic texts.

Keywords: corpus-driven research, intra-speaker variation, lexical stress, poetic language, variation in phonology

ВНУТРИИДИОЛЕКТНАЯ ВАРИАТИВНОСТЬ УДАРЕНИЯ В РУССКОМ ЯЗЫКЕ: КОРПУСНОЕ ИССЛЕДОВАНИЕ НА МАТЕРИАЛЕ РУССКОЙ ПОЭЗИИ

Пиперски А. Ч. (apiperski@gmail.com)

Российский государственный гуманитарный университет; Национальный исследовательский университет Высшая школа экономики, Москва, Россия

Кухто А. В. (anton.kukhto@gmail.com)

Московский государственный университет имени М. В. Ломоносова, Москва, Россия

Русское ударение обнаруживает вариативность не только у разных носителей, различающихся между собой по региональной принадлежности, социальному статусу, возрасту и т. д., но и внутри одного идиолекта. Вариативность второго рода трудно исследовать, поскольку для этого необходимы большие устные корпуса текстов от одного носителя. Их, однако, можно заменить поэтическими корпусами. В статье мы автоматически анализируем тексты десяти поэтов, взятые из Национального корпуса русского языка, чтобы найти словоформы с акцентной вариативностью. Число таких форм у одного носителя лежит в интервале от 30 до 200 словоформ разных частей речи. В статье предлагается количественная мера для оценки общей вариативности ударения, не зависящая от размера корпуса; её сравнение для разных авторов показывает, что вариативность снижается со временем, по меньшей мере в поэтических текстах.

Ключевые слова: вариативность в фонологии, внутриидиолектная вариативность, корпусное исследование, словесное ударение, язык поэзии

1. Introduction

Word stress in Russian is generally assumed to be stored lexically and to be driven by morphology (Zalizniak 1985; Knyazev and Pozharitskaya 2012), with some default rules also present in phonology (Lavitskaya and Kabak 2014). This means that in each cell of a given word's paradigm the stress is fixed on a certain syllable, depending on the properties of the word's morphemes. Nevertheless, there exists a substantial amount of variation even within Standard Russian as well as across different regional varieties. Such variation is reflected in Russian pronouncing dictionaries, and some of the words with variable stress tend to become especially prominent among the Russian public. For example, consider the controversy concerning the pronunciation of the word *'zvonit* or *zvo'nit* 'he / she / it calls' where initial stress is heavily stigmatized. Despite this fact, many aspects of stress variation in Russian remain, to the best of our knowledge, understudied, research in this field being rather scarce. It is especially true of intra-speaker variation, as opposed to inter-speaker variation (for the delimitation of these types, cf. Honeybone 2011). Whereas stylistic, regional and chronological inter-speaker variation has received some scholarly attention (Lagerberg 2011; Lehfeldt 2014), intra-speaker variation has not.

Approaching the problem of intra-speaker stress variation, we can take at least two paths of its examination. One is to analyze it through recourse to experimental methods. Such a study focusing on intra-speaker variation was reported by Knyazev, Kukhto, and Piperski (2015) and by Kukhto and Piperski (2016). Speakers of Russian were requested to read out loud sentences containing the word forms *prodal* 'he sold' and *obnjal* 'he hugged'. It was found that the position of stress in such verbal forms is at least partially determined by the stress of the immediately following direct object, making initial stress *'prodal* in sequences like *prodal 'daču* 'he sold a cottage' more likely than in sequences like *prodal bra'slet* 'he sold a bracelet', and the other way round

with *pro'dal*. That is to say, regarding the results of this experiment, lexical stress shows a tendency to adhere to the Principle of Rhythmic Alternation (see Schlüter 2015), thus exhibiting a general preference for the alternation of stressed and unstressed syllables.

A drawback of this type of experimental study is that it only allows us to examine variation within a limited set of words, primarily because of the restricted resources in terms of speakers' patience (with lexical stress there is an additional problem, namely the speakers' tendency to guess the purpose of the experiment at some point, which renders further investigation useless). Another limitation is that an approach of this type cannot give us any information about real-time—as opposed to apparent-time—change, unless a series of experiments is performed repeatedly over a number of years. Therefore, a corpus study may better help to observe the full scale of variation.

At first sight, spoken corpora appear to be an effective solution. However, most of them are better suited for analyzing inter-speaker variation rather that intra-speaker variation, since large corpora of text produced by one and the same speaker (and annotated for lexical stress with reasonable quality) remain a desideratum. Yet, there is one straw to be grasped, and that is poetic corpora.

Russian syllabotonic poetry has been heavily reliant on lexical stress since the mid-18th century (Gasparov 2000). As shown by Kolmogorov and Prokhorov (1968), a fundamental principle of Russian classical poetry is that the lexical stresses of polysyllabic words can only occupy strong metrical positions, which, in reverse, makes poetry a valuable resource for studying the stress in such words. One might object that stress in poetic texts does not directly reflect the stress in prose; however, the role of poetic license in Russian is often exaggerated. It is true that stress in poetic texts often deviates from the norms of Modern Standard Russian, but such deviations are not random and rarely bend actual stresses simply to fit the meter. To quote Bulakhovskij (1952: 22), "there is a wide-spread wrong opinion that requires some comments. Many people believe that poets freely distort stress patterns in order to fit the rhythm. However, no cultivated poet ever allows himself more variation than is actually present in the standard language of his time".[1] Further evidence for the non-existence of arbitrary poetic license can be found in Gorbachevich (1989: 77–8).

2. Data

In the present paper we analyze intra-speaker variation in the texts of 10 Russian poets: Alexander Pushkin (1799–1837), Nikolay Yazykov (1803–1846), Mikhail Lermontov (1814–1841), Apollon Maykov (1821–1897), Vyacheslav Ivanov (1866–1949), Mikhail Kuzmin (1872–1936), Nikolay Gumilev (1886–1921), Aleksandr Tvardovsky

[1] «Особых замечаний требует один широко распространенный предрассудок. Многие думают, будто поэты по требованию ритма разрешают себе вольное обращение с ударением, доходящее иногда до искажений. На самом деле ни один культурный поэт никогда не позволял себе и не позволяет колебаний больших, чем те, которые реально существуют в литературном употреблении его времени» (Bulakhovskij 1952: 22).

(1910–1971), Konstantin Simonov (1915–1979), and David Samoylov (1920–1990).[2] For these poets, we considered all texts from the poetic subcorpus of the Russian National Corpus (RNC) that are marked as purely syllabotonic, i.e. trochaic, iambic, dactylic, amphibrachic or anapaestic (this decision was made to facilitate the judgments about stress placement and, indeed, to make them possible). The size of the corpus for each poet is given in Table 1:

Table 1. Corpora sizes for 10 poets

Poet	Texts	Tokens	Word types
Pushkin	855	182,014	35,377
Yazykov	354	59,008	15,904
Lermontov	441	125,883	23,122
Maykov	553	107,696	26,196
Ivanov	1,025	103,357	28,717
Kuzmin	553	57,745	17,860
Gumilev	446	57,390	17,245
Tvardovsky	306	101,448	21,272
Simonov	204	51,332	14,505
Samoylov	751	58,179	18,907

Using the stress annotation provided by the RNC, we generated full lists of word forms that occur with varying stress placement, e.g., *glu'boko* and *glubo'ko* 'deeply', in texts by the same poet. This means the study we conducted was corpus-driven rather than corpus-based, because this method of analysis relies heavily on the automatic processing of digitized texts and would be impossible to implement manually. However, the results still had to be filtered manually, since there are many homographic but not homophonous forms (such as *'bedy* 'troubles (NOM/ACC.PL)' or *be'dy* 'of the trouble (GEN. SG)'), as well as some mistakes in the RNC markup. For some word forms, it is difficult to make a clear distinction between stress variation and polysemy / homography of different words or different forms of the same word. For instance, this is the case with *'devica* and *de'vica* 'maid', where there are subtle differences in meaning, or in the case of short forms of adjectives used attributively and predicatively (*'mračna noč* 'gloomy night' vs. *noč mrač'na* 'the night is gloomy', cf. Kuleva 2008: 10). In each case a separate decision had to be made based on the judgments of the authors of the present paper (both trained linguists and speakers of Modern Standard Russian).

It should be noted that we make the simplifying assumption that a speaker is not subject to language change throughout their life. We adhere to the apparent-time hypothesis in its strict form (Milroy & Gordon 2003: 35–7), claiming that an individual's language remains stable after being acquired in childhood. This simplification is necessary because some word forms with variable stress are attested only a small number of times, and we cannot be certain whether they actually reflect intra-speaker variation or whether they have undergone intra-speaker change over time.

[2] These authors were chosen to ensure balance with respect to chronology and corpus size.

The distribution of word forms with variable stress across different parts of speech for the 10 poets is shown in Table 2[3]:

Table 2. Word forms with variable stress according to part of speech

Poet	Nouns	Adjectives	Adverbs	Verbs	Numerals	Total
Pushkin	65 (35%)	38 (20%)	8 (4%)	74 (40%)	2 (1%)	187
Yazykov	14 (34%)	18 (44%)	4 (10%)	5 (12%)	0 (0%)	41
Lermontov	50 (41%)	27 (22%)	8 (7%)	36 (30%)	0 (0%)	120
Maykov	52 (39%)	22 (16%)	13 (10%)	48 (36%)	0 (0%)	135
Ivanov	68 (57%)	25 (20%)	3 (2%)	26 (21%)	0 (0%)	122
Kuzmin	14 (22%)	13 (21%)	12 (19%)	24 (38%)	0 (0%)	63
Gumilev	22 (36%)	17 (28%)	6 (10%)	16 (26%)	0 (0%)	61
Tvardovsky	13 (30%)	5 (12%)	10 (23%)	15 (35%)	0 (0%)	43
Simonov	8 (27%)	6 (20%)	7 (23%)	9 (30%)	0 (0%)	30
Samoylov	6 (40%)	4 (27%)	3 (20%)	2 (13%)	0 (0%)	15

As an example, below is a full list of forms with variable stress attested in the corpus of syllabotonic poetry by Aleksandr Tvardovsky, grouped by parts of speech, and their frequencies:

Nouns:		
'vësnu × 1	*ve'snu* × 2	‘spring (ACC.SG)’
vo'rota × 5	*voro'ta* × 3	‘gate’
'kladbišče × 3	*klad'bišce* × 1	‘graveyard’
'krovi × 4	*kro'vi* × 1	‘blood (GEN.SG)’
'mosta × 4	*mo'sta* × 15	‘bridge (GEN.SG)’
'nuždy × 6	*nu'ždy* × 9	‘need (GEN.SG)’
'okon × 1	*o'kon* × 2	‘window (GEN.PL)’
'poldni × 1	*pol'dni* × 1	‘noon (NOM.PL)’
'polnoči × 2	*pol'noči* × 1	‘midnight (GEN.SG)’
'stenax × 1	*ste'nax* × 2	‘wall (LOC.PL)’
'sudeb × 5	*su'deb* × 2	‘fate (GEN.PL)’
'utra × 1	*u'tra* × 13	‘morning (GEN.SG)’
'hody × 2	*ho'dy* × 1	‘pathway (NOM.PL)’

[3] It must be noted that we deal with word forms rather than lemmas, since it is often the case that some forms of a word exhibit variation, while others do not. This implies that some variation remains unnoticed. For instance, if a speaker has variation in the word form *'stenax ~ ste'nax* ‘wall (LOC.PL)’, it is likely to also exist in DAT.PL and INS.PL; however, these case forms may be unattested.

Adjectives:		
'blizki × 2	*bli'zki* × 1	'close (PL)'
'bosye × 1	*bo'sye* × 1	'barefoot (PL)'
'davnišnjaja × 1	*da'vnišnjaja* × 1	'bygone (F.SG)'
'suhi × 2	*su'hi* × 1	'dry (PL)'
'ščastliv × 10	*šča'stliv* × 1	'happy (M.SG)'

Adverbs:		
vy'soko × 4	*vyso'ko* × 8	'high'
glu'boko × 6	*glubo'ko* × 6	'deep'
da'lëko × 23	*dale'ko* × 12	'faraway'
'zadolgo × 1	*za'dolgo* × 1	'long before'
izda'lëka × 3	*izdale'ka* × 17	'from afar'
'mel'kom × 1	*mel"kom* × 1	'swiftly'
'navek × 1	*na'vek* × 14	'forever'
'naverx × 1	*na'verx* × 1	'upwards'
'poverx × 1	*po'verx* × 1	'on top of'
'totčas × 13	*tot'čas* × 11	'immediately'

Verbs:		
'valit × 1	*va'lit* × 3	'make fall (3SG.PRES)'
'vzjalsja × 3	*vzjal'sja* × 1	'undertake (M.SG.PST)'
'drožit × 1	*dro'žit* × 6	'tremble (3SG.PRES)'
za'lilsja × 1	*zalil'sja* × 1	'burst into (M.SG.PST)'
'minulo × 3	*mi'nulo* × 2	'pass (N.SG.PST)'
ne'obžitoj × 1	*neob'žitoj* × 1	'not render habitable (NEG.PTCP.PASS.PST)'
'obžitoj × 1	*ob'žitoj* × 1	'render habitable (PTCP.PASS.PST)'
'obnjal × 2	*ob'njal* × 2	'hug (M.SG.PST)'
'podnjav × 2	*po'dnjav* × 2	'raise (GER.PFV)'
po'dnjalsja × 8	*podnjal'sja* × 1	'raise oneself (M.SG.PST)'
ro'dilsja × 8	*rodil'sja* × 1	'be born (M.SG.PST)'
so'brala × 2	*sobra'la* × 1	'gather (F.SG.PST)'
so'bralsja × 7	*sobra'lsja* × 2	'set out (M.SG.PST)'
u'dalsja × 2	*uda'lsja* × 1	'succeed (M.SG.PST)'
u'përšis' × 2	*uper'šis'* × 1	'lean against (GER.PFV)'

A look at Tables 1 and 2 alone does not make it possible to compare the amount of variation among individual poets. It is clear that the count of tokens with variable stress depends heavily on the corpus size and the frequency distribution of words within it. A larger corpus is likely to provide more opportunities for a token with variation to surface and for variation to come to light. Type-to-token ratio (TTR) also plays a role, since a corpus with a low TTR includes only a small number of types, which brings down the number of types with variation, even though it is more likely

to be attested for each of them. Alternatively, a corpus with a high TTR contains many words, but these appear only a few times each, which means variation is also likely not to surface. For this reason, we need to reduce the counts to find an interpretable measure of variation that is independent of corpus size.

3. Model

In this section, we present a simplified model of our data. Let us assume that there are two types of words—those with variable and those with invariable stress. Let us further assume that any word with variable stress has two possible stresses, one of them surfacing with a probability of 0.25, the other with a probability of 0.75.[4] The probability of variation being attested (a_n) then depends on the number of occurrences of a word in a corpus (n). It is equal to $1 - (0.25^n + 0.75^n)$, where n is the number of its attestations and 0.25^n and 0.75^n are the probabilities of an underlyingly variable word to surface in one of its two forms at all times. For instance, an underlyingly variable word occurring only once cannot exhibit any variation ($a_1 = 0$), an underlyingly variable word occurring twice will exhibit variation with a probability of $a_2 = 0.375$, an underlyingly variable word occurring three times will exhibit variation with a probability of $a_3 = 0.562$, etc.

Let us make an additional assumption that a word w belongs to the class of underlyingly variable words with a probability of v, and to the class of underlyingly invariable words with a probability of $1 - v$. Thus, the probability of a word occurring $f(w)$ times to exhibit variation is equal to $v \times a_{f(w)}$, and the expected number of words with stress variation in the corpus equals $\sum v \times a_{f(w)}$ over all w's.

This logic can easily be reversed. Once we know the number of words K with stress variation in the corpus of a poet, we can estimate the value of v so that it would yield the same expected count of words with variation: $\hat{v} = K / \sum a_{f(w)}$. This value seems to be a good estimate of how much intra-speaker variation a given speaker has. However, it is still not robust against corpus size. This can be seen in Table 3, where $\hat{v}$ was calculated for 10 sizes of subcorpora of Pushkin's texts, ranging from 10% to 100% of the total amount of texts available (for each subcorpus size from 10% to 90%, a random selection of texts was taken 20 times, and the mean value of $\hat{v}$ for these 20 trials is provided):

Table 3. The value of $\hat{v}$ in the subcorpora of Pushkin's texts of different sizes

	10%	20%	30%	40%	50%
$\hat{v}$	0.0135	0.0152	0.0162	0.0168	0.0172
	60%	70%	80%	90%	100%
$\hat{v}$	0.0177	0.0185	0.0188	0.0191	0.0194

[4] 1:3 (0.25:0.75) is the average distribution of the two variants for words with variable stress that were attested at least 5 times in the corpus of a single author. A more elaborate model might take into account that different words have different frequencies of variant forms, e.g., by looking at these frequencies in the decade immediately following the poet's birth, but it remains to be tested whether such a model would fit the data better.

As can clearly be seen, the estimated variability is higher in larger corpora. This might be due to the fact that larger corpora contain a larger amount of infrequent word forms, where the speakers are less certain about the stress. In order to embed this in our model and render the measure of variability more robust, let us make one more assumption: the probability of having variable stress is not equal to v for all words, but also depends on the rank of the word in the frequency list. The higher the word is in the frequency list, the less variability one would expect, and vice versa.

Let us suppose that the probability that a word belongs to the class of underlyingly variable words is not v, but $v \times r^s(w)$, where $r(w)$ is the rank of the word on the frequency list ($r(w) = 1$ for the most frequent word, etc.), and s is a constant. If s equals 0, it brings us to our starting point (the probability of a word being underlyingly variable is always equal to v); however, if s is a small fraction above zero, it makes less frequent words more variable.

After testing all possible values of s between 0 and 0.30 with a step of 0.001 on the Pushkin and Tvardovsky corpora, we arrived at the conclusion that $s = 0.20$ makes the estimation of v least dependent on corpus size (namely, the standard deviation of the mean estimated values of v for the 10%-samples, 20%-samples, …, 90%-samples, as well as for the entire corpus is smallest when compared to these values). This makes the final version of our model look as follows:

A word w with the rank $r(w)$ on the frequency list attested n times has underlyingly variable stress (= two variant forms distributed as 1:3) with probability $p = v \times r^{0.2}(w)$.[5] If it has underlyingly variable stress, it surfaces with different stresses with probability $a_n = 1 - (0.25^n + 0.75^n)$. Thus, a word is likely to be attested with variable stress with a probability $P = v \times r^{0.2}(w) \times a_n$. A sum of P's for all words ($\sum P = \sum v \times r^{0.2}(w) \times a_n$) yields an expected number of words with variation K. Once we know the value of K, we can obtain an estimate of v:

$$\hat{v} = \sum (r^{0.2}(w) \times a_n) / K$$

$\hat{v}$ then reflects how much intra-speaker variability a speaker has.

The values of $\hat{v}$ for the 10 poets studied are quoted in Table 4:

Table 4. Estimates of stress variability for the 10 studied poets

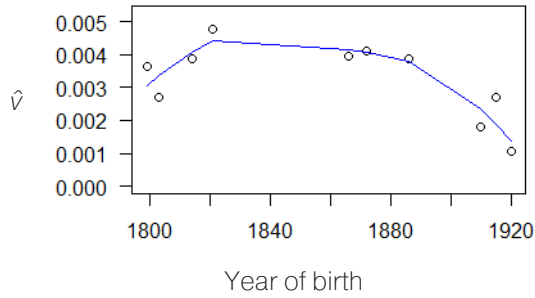
Poet	$\hat{v}$
Pushkin	0.00363
Yazykov	0.00268
Lermontov	0.00387
Maykov	0.00475
Ivanov	0.00394

Poet	$\hat{v}$
Kuzmin	0.00412
Gumilev	0.00387
Tvardovsky	0.00180
Simonov	0.00270
Samoylov	0.00106

[5] Strictly speaking, we should not allow this value to exceed 1, making it $p = \min(v \times r^{0.2}(w), 1)$. However, this is not necessary in practice, since v is usually small and $r(w)$ is not greater than 40,000 for any of our corpora.

It must be noted that there are some simplifications inherent to our model. For instance, it does not take into account the fact that there are non-syllabic, monosyllabic and clitic words that cannot have any stress variation at all. Nonetheless, these words are generally in the top part of the frequency list, which means that the model correctly assigns them a smaller probability of variation simply because their r (w) is smaller.

Anyway, the value of $\hat{v}$ is a good estimator of how much variation a speaker has, and these values can be compared across speakers, in spite of differences in corpus size. The comparison of poets listed in Table 4 shows that the amount of variation was generally higher in the 19th century and gradually diminished later in the 20th century. This is visualized in Graph 1, where the estimated values of $\hat{v}$ are plotted against the birth years of individual poets (a LOWESS curve is added to the plot to make it more illustrative):



Graph 1. Poets' year of birth and $\hat{v}$

The conclusion we have arrived at using our apparent-time model can be compared to parallel measurements of inter-speaker variation within chronological layers. For these, we created four corpora with texts from four decades (1801–1810, 1851–1860, 1901–1910, 1951–1960), each comprising a random selection of poems totaling up to 100,000 word forms from all poets active during a given decade. Their analysis employing the same methods as above also shows that the rate of variation tends to diminish in bulk towards the middle of the 20th century. The results are presented in Table 5.

Table 5. Stress variation in real time by decades

Decade	Word types	Words forms with variable stress	$\hat{v}$
1801–1810	26,099	187	0.00552
1851–1860	26,215	89	0.00266
1901–1910	27,100	106	0.00299
1951–1960	29,729	59	0.00175

In terms of interpretation, these results can likely be explained by the fact that the norms of Standard Russian were becoming more rigid towards the 20th century. However, even rigid norms turn out not to be powerful enough to dispose of stress variation as a whole.

4. Conclusions

In our paper, we discussed variation in Russian stress. It manifests itself not only as inter-speaker variation depending on regional variety, social status, age, etc., but also as intra-speaker variation, which is difficult to capture. Using the evidence of Russian poetry from the 19th and 20th century in order to study intra-speaker variation, we propose a quantitative measure of overall stress variability independent of corpus size and show that there is a tendency for variation to diminish over time. To provide further support for these findings, we are planning to enlarge the analyzed corpus in future studies.

Acknowledgements

We are grateful to Michael Daniel, Boris Iomdin and Georgy Moroz for their valuable comments, corrections and suggestions.

References

1. *Bulakhovskij L. A.* (1952), A course in Standard Russian [Kurs sovremennogo russkogo jazyka], Kiev, Radjans'ka škola. Vol. 1.
2. *Gasparov M. L.* (2000), A concise history of Russian verse [Očerk istorii russkogo stiha], Moscow, Fortuna Limited.
3. *Gorbachevich K. S.* (1989), The norms of Modern Standard Russian [Normy sovremennogo russkogo literaturnogo jazyka], Moscow, Prosveščenie.
4. *Honeybone P.* (2011), Variation and linguistic theory, in Maguire W., McMahon A. (eds.), Analysing variation in English, Cambridge: Cambridge University Press, pp. 151–77.
5. *Knyazev S., Kukhto A., Piperski A.* (2015), Postlexical stress in Standard Russian: A case study, Poster presented at the Phonetics and Phonology in Europe conference, University of Cambridge, 28–30 June 2015.
6. *Knyazev S. V., Pozharitskaya S. K.* (2012), Modern Standard Russian: Phonetics, Writing System, Spelling, Orthoepy [Sovremenny russkij literaturnyj jazyk: fonetika, grafika, orfografija, orfoèpija], Moscow, Akademičeskij proekt.
7. *Kolmogorov A. N., Prokhorov A. V.* (1968), Toward the foundation of Russian classical metrics [K osnovam russkoj klassičeskoj metriki], in The commonwealth of sciences and the secrets of creativity [Sodružestvo nauk i tajny tvorčestva], pp. 397–432.
8. *Kukhto A., Piperski A.* (2016), Free stress variation and the rise of rhythmic rules in Russian, Paper presented at the Workshop on Variation in Phonology at the 13th Old World Conference in Phonology, Eötvös Loránd University, Budapest, 13–16 January 2015.
9. *Kuleva A. S.* (2008), Apocopated adjectives in the language of Russian poetry [Usečennye prilagatel'nye v jazyke russkoj poèzii], PhD thesis summary [Avtoref. diss. … kand filol. nauk], Moscow.

10. *Lagerberg R.* (2011), Variation and Frequency in Russian Word Stress. Slavistische Beiträge, 477. Munich and Berlin: Verlag Otto Sagner, 2011.
11. *Lavitskaya Y., Kabak B.* (2014), Phonological default in the lexical stress system of Russian: Evidence from noun declension, Lingua, Vol. 150, October 2014, pp. 363–85.
12. *Lehfeldt W.* (2014), Accent and stress in Modern Russian [Akcent i udarenie v sovremennom russkom jazyke], Moscow: Jazyki slavjanskoj kul'tury.
13. *Milroy L., Gordon M.* (2003), Sociolinguistics: Method and interpretation, Malden: Blackwell.
14. *Russian National Corpus* (RNC), www.ruscorpora.ru
15. *Schlüter J.* (2015), Rhythmic influence on grammar: Scope and limitations, in Vogel R., van de Vijver R. (eds.), Rhythm in cognition and grammar: A Germanic perspective, Berlin, de Gruyter Mouton, pp. 179–206.
16. *Zalizniak A. A.* (1985), From Proto-Slavic to Russian accentuation [Ot praslavjanskoj akcentuacii k russkoj], Moscow, Nauka.

«НО ПО РАСЧЕТУ ПО МОЕМУ ДОЛЖНА РОДИТЬ»: КОНСТРУКЦИИ С СОЮЗОМ *НО* ПО ДАННЫМ КОРПУСОВ С ПРОСОДИЧЕСКОЙ РАЗМЕТКОЙ

Подлесская В. И. (podlesskaya@ocrus.ru)

Российский государственный гуманитарный университет;
Российская академия народного хозяйства
и государственной службы, Москва, Россия

Ключевые слова: сочинение, сложное предложение, русский язык, корпус, устная речь, просодия

"NO PO RASCHOTU PO MOEMU DOLZHNA RODIT": THE RUSSIAN CONJUNCTION *NO* VIEWED THROUGH THE PRISM OF PROSODICALLY ANNOTATED CORPUS DATA

Podlesskaya V. I. (podlesskaya@ocrus.ru)

Russian State University for the Humanities;
Russian Academy of National Economy and Public Administration, Moscow, Russia

The paper focuses on Russian coordinate construction with clauses (or VPs) combined by means of the adversative conjunction NO. Prosodically, the construction may come up in two forms: (a) as a single illocution with the first clause pronounced with a rising pitch that projects discourse continuation, and (b) as two separate illocutions with the first clause pronounced with a falling pitch that projects no continuation. Basing on the data from the Prosodically Annotated Corpus of Spoken Russian, prosody and grammar of (a) and (b) were analyzed qualitatively and quantitatively. Type (b) appeared to be less frequent (comprising, approximately, 30 % of the total number of occurrences) and systematically favored in contexts where the second clause is complicated by a "heavy" topical constituent.

Key words: coordination, clause combining, Russian, corpus, natural discourse, prosody

1. Постановка вопроса

В русском языке в прототипических бинарных противительных конструкциях с союзом НО, см. (1), союз просодически примыкает ко второму компоненту, при этом первый компонент обычно бывает цельнооформленным с точки зрения лексики и грамматики[1]:

(1) *Она не родила, но по расчету по моему: должна родить.*

Иначе говоря, при развертывании дискурса во времени слушающий по получении первого компонента (*она не родила*) не располагает никакими конструкционными или лексическими сигналами, которые бы предсказывали бы тот или иной маршрут последующего развертывания дискурса. Напротив, второй компонент содержит элементы, которые позволяют привязать его к уже артикулированному сегменту: это, прежде всего, союз НО, связывающий второй компонент противительным отношением (плюс, в данном конкретном случае, нулевое подлежащее второго компонента, привязывающее анафорическим отношением подлежащее первого компонента). Если использовать терминологию, принятую в рамках анализа бытового диалога, можно говорить о том, что первый компонент «не проецирует» продолжения, т.е. не содержит сегментных или конструкционных «проекторов» (Auer 2002). Этим, в частности, противительные конструкции отличаются от конструкций с препозитивным уступительным протазисом, см. (1'):

(1') *хотя она не родила…*

в которых проектором является уступительный союз, уверенно предсказывающий появление аподозиса.

Однако в устной речи, помимо лексических и конструкционных проекторов, важнейшую роль в организации дискурса играют проекторы просодические. Так, устная версия примера (1), в общем случае, реализуется с подъемом тона в главном фразовом акценте первого компонента, который располагается на ударном слоге слова *родила*, этот подъем является стандартным просодическим средством оформления дискурсивной незавершенности в русском языке. Иначе говоря, реализация «с подъемом» является стандартным проектором, сигнализирующим слушающему, что следует ожидать продолжения. В письменной речи этот проектор конвенционально нотируется с помощью пунктуации, ср. запятую перед НО в примере (1). Заметим, что просодическая незавершенность первого компонента является лишь одной из опций при произнесении (1) — первый компонент можно произнести и с падением тона в главном фразовом акценте, что на письме конвенционально передается с помощью точки перед НО:

(1'') *Она не родила. Но по расчету…*

[1] Исследование поддержано РФФИ (грант16-06-00226)

В этом случае первый компонент оказывается завершенным не только лексико-грамматически, но и просодически, т. е. по его завершении у слушающего — в отсутствие каких бы то ни было проекторов — не формируется никаких прогнозов относительно дальнейшего развертывания дискурса.

В данной работе, опираясь на корпусные данные, я попытаюсь ответить на следующие вопросы:

- Каков арсенал просодических конфигураций, используемых в составе конструкций с НО?
- Какие из этих конфигураций могут использоваться в качестве проекторов в разворачивающейся структуре дискурса?
- Как просодические проекторы могут согласовываться (или не согласовываться!) с лексико-грамматическими?
- Какие уроки это позволяет извлечь для понимания общих процессов порождения и понимания речи?

Я использую материал трех корпусов электронной коллекции «Рассказы о сновидениях и другие корпуса звучащей речи» (Prosodically Annotated Corpus of Spoken Russian, PrACS-Russ), содержащей аудиофайлы с синхронизированными просодически размеченными транскриптами (Spokencorpora 2013). В этих корпусах были проанализированы все имеющиеся вхождения союза НО:

- «Рассказы сибиряков о жизни» (17 монологов, респонденты от 19 до 70 лет, около 5000 словоупотреблений), 21 вхождение союза НО, примеры с индексом Sib;
- «Веселые истории из жизни» (40 монологов, респонденты от 18 до 60 лет, около 10000 словоупотреблений, плюс письменные версии этих же рассказов, около 7000 словоупотреблений), 53 вхождения союза НО в устной части, примеры с индексом FS-sp, 31 вхождение — в письменной части, примеры с индексом FS-wr;
- «Истории о подарках и катании на лыжах» (20 рассказов по картинкам и 20 пересказов тех же сюжетов по памяти, респонденты от 20 до 30 лет, около 4500 словоупотреблений), 32 вхождения союза НО, примеры с индексами Ski, Pr.

Дальнейшее изложение будет строиться следующим образом. В разделе 2 я проанализирую наблюдаемые в корпусе просодические явления, которые позволяют интерпретировать фрагмент перед НО как дискурсивно завершенный или, напротив, — как проецирующий продолжение, и продемонстрирую случаи, которые не поддаются однозначной интерпретации. В разделе 3 обсуждаются лексико-грамматические явления, которые устойчиво кластеризуются с просодической завершенностью или с просодической незавершенностью. В разделе 4 я приведу количественные данные и, в частности, сравню распределение просодической завершенности/незавершенности с распределением «точка/запятая» перед НО в письменной речи. В завершение я постараюсь сформулировать некоторые обобщения.

2. Союз НО: после просодически завершенных или после просодически незавершенных фрагментов?

2.1. Очевидные случаи

Как и следовало ожидать, в нашем материале широко представлены прототипические бинарные противительные конструкции, в которых компоненты иллокутивно однородны, т. е. имеют одинаковую иллокутивную силу (преимущественно, сообщение, так как мы имеем дело с коллекцией нарративов) и связаны противительным отношением (о семантике противительности в связи со значением союза НО, см. в частности, Санников 1989, Malchukov 2004). Первый компонент таких конструкций может быть просодически завершенным, т. е. не содержать просодического проектора, или, напротив, просодически незавершенным, что проецирует продолжение, т. е. формирует у слушающего ожидание продолжения. Просодическая завершенность перед НО стандартно маркируется падением тона в главном фразовом акценте первого компонента по типу ИК-1 в терминологии интонационных конструкций (Брызгунова 1882, Янко 2008). Так, в следующем фрагменте все четыре вхождения союза следуют за просодически завершенными фрагментами[2]:

(2) FS_02-f_Sp

7.	И-и /решили \/–мы ..(0.39) мм(0.28) /поехать на смотровую \площ<u>а</u>дку[w].
8.	..(0.31) {ЧМОКАНЬЕ 0.14} Но там \н<u>е</u> было парковочных мест.
9.	/Мы решили заехать на /территорию \университ<u>е</u>та[w].
10.	..(0.31) Но /там висит ↑\–**кирпи-и**<u>**ч**</u>.
11.	....(1.09) \В<u>о</u>-от-т.
12.	Но /так как я всё время /наруш<u>а</u>ю,
13.	то /мы решили
14.	что под /кирпич **м**<u>**о**</u>**жно** прое-ехать.
15.	А /денег у нас всё вре= || всё || всего осталось /сто \рубл<u>е</u>й.
16.	..(0.24) /Но решили поехать под /кирпи<u>ч</u>,
17.	потому что там \вс<u>е</u> ездят.

Просодическая незавершенность перед НО маркируется чаще всего (детальный количественный анализ отложим до раздела 4) «прототипическим русским подъемом» с падением на заударных, если они есть, т. е. по типу ИК-3 (Янко 2008:31):

---

[2] Об используемой системе дискурсивной транскрипции см. Кибрик, Подлесская (ред.) 2009, Spokencorpora 2013. Индекс примера содержит отсылку к одному из трех обследованных корпусов и ярлыку текста в составе корпуса.

(3) 04-m_Pr-T

13.	….(1.44) И-и узнав /отв<u>е</u>т^{h},
14.	…(0.77) он очень /расстр<u>о</u>ился,
15.	…(0.70) но /всё-таки придумал выход из \полож<u>е</u>ния.

В более редких случаях незавершенность перед НО маркируется конфигурацией типа ИК-4 — падением на ударном слоге с последующим подъемом на заударных или непосредственно на ударном слоге, если заударных нет. Эта конфигурации в русском языке, по наблюдениям Т. Е. Янко (2008: 33, 200–225), связана со значением «рассказа по порядку», а также с сопоставлением и противопоставлением:

(4) 08-f_Pr-T

32.	…(0.89) /Папа \↑отказ<u>а</u>лся ↑тактично^{w},
33.	..(0.47) /но не \→растер<u>я</u>лся^{w}:

К числу очевидных случаев можно отнести и такие, в которых между компонентами, связанными противительным отношением, имеется вставка, обычно уточняющая значение первого компонента. Эта вставка часто произносится как парентеза — в более узком частотном диапазоне, со сниженным уровнем громкости — и может завершаться нисходящим акцентом, но семантически сферой действия просодической незавершенности является не материал вставки, а именно фрагмент, вводимый союзом НО. Так, в следующем примере просодическим проектором является восходящий акцент в строке 28, именно он предсказывает появление строки 31, связанной противительным отношением со строкой 28, а строки 29–30 просодически реализуются как парентеза:

(5) Sib_08-f

28.	…(0.59) Как-то /Никита /гонялся за ним по всей /квартире,
29.	..(0.43) потому что он ..(0.45) \сожрал /б<u>о</u>чечку от лото,
30.	\ун<u>ё</u>с в пасти,
31.	но не \см<u>о</u>г догнать.

Во всех перечисленных случаях просодическая завершенность \ незавершенность фрагмента, предшествующего союзу НО, интерпретируется слушающим однозначно. Вместе с тем, выясняется, что просодическое оформление НО-конструкций в нашем материале далеко не исчерпывается продемонстрированными выше «очевидными» случаями.

2.2. Проблемные случаи

2.2.1. Первую группу проблемных случаев составляют контексты, в которых фрагмент, предшествующий НО, маркирован акцентом типа ИК-6 по Е. А. Брызгуновой — с подъемом на ударном слоге, за которым не следует падения на заударных. Эта конфигурация, которую условно можно назвать интонацией многоточия, особенно при растянутом ударном слоге, выражает, согласно Т. Е. Янко, значение имитации ментальной деятельности (припоминание, недоумение), однако при этом широко используется и для выражения незавершенности при описании череды событий, открытого списка (Янко 2008: 109–113, 166–167). Получая на вход фрагмент, оформленный таким образом, слушающий допускает, что возможно продолжение, но оно жестко не проецируется. Так, в следующем примере, строки 11–13 представляют собой незаконченное перечисление попыток выбрать подарок, главные фразовые акценты в этих строках реализуются как ИК-6 с растянутой ударной гласной. В принципе, ни лексико-грамматическое, ни просодическое оформление этих строк не проецируют их потенциальную связь со следующим фрагментом, однако и не противоречат такой связи. В той системе нотации, которая используется в корпусах Spokencorpora 2013, такого рода омонимия дискурсивного статуса делает допустимыми два варианта нотации в позиции перед НО, ср. (6a) — знак «…» плюс заглавная буква, т. е. завершенность перед НО, иллокутивная граница:

(6a) 01-f_Pr-R

11.	..(0.32) /То он хотел купить /\су-умку,,,
12.	то он хотел купить ..(0.17) /\часы-ы,,,
13.	то /\манеке-ен…
14.	..(0.35) Но /всё у него не \получалось.

и (6b) — знак «,,,» (ослабленное, внутрииллокутивное многоточие) плюс строчная буква, т. е. незавершенность перед НО в составе единой иллокуции:

(6b) 01-f_Pr-R

11.	..(0.32) /То он хотел купить /\су-умку,,,
12.	то он хотел купить ..(0.17) /\часы-ы,,,
13.	то /\манеке-ен,,,
14.	..(0.35) но /всё у него не \получалось.

Можно заключить, что в такого рода контекстах противопоставление по просодической завершенности оказывается нейтрализованным.

2.2.2. О нейтрализации противопоставления по просодической завершенности/незавершенности можно говорить и в тех случаях, когда фрагмент перед НО маркируется нисходящим акцентом по типу ИК-2 (интенсивное падение с большим

перепадом частот). Этот тип акцента может использоваться для выражения контраста (Янко 2008: 32–38) и как таковой оказывается адекватным в контексте противительности. Так, в следующем примере по типу ИК2 реализуется главный фразовый акцент в строке 85 на слове *пиво* (строка 86 — парентетический дискурсивный маркер). Эта строка звучит также, как могло бы звучать изолированное верификативное эмфатически окрашенное высказывание, например, с частцей же: *Он же выпил одно пиво!* (= ‘как могло оказаться, что он пьян?!’)

(7) FS_18-f_Sp

84.	у меня вот такое /ощущ<u>е</u>ние,
85.	что он выпил просто одно **пи<u>в</u>о**,
86.	не \зн<u>а</u>ю,
87.	но был **жутко** совершенно какой-то .. невообразимо \пь<u>я</u>ный.

Аналогичным образом, верификативное контрастивное значение вносит акцент по типу ИК-2 на слове *купил* в строке 35 в следующем примере (= ‘он ее-таки купил!’):

(8) 01-f_Pr-T

35.	В общем в /результате */он конечно машину \куп<u>и</u>л*,
36.	но она была /\ма-аненькая-\м<u>а</u>ненькая!,

В примерах (7), (8) нотация с точкой перед НО и заглавной буквой в транскрипционной системе Spokencorpora 2013 тоже правомерна. В такого рода случаях, как и в случаях с интонацией многоточия, просодическая реализация удачно вписывается в конструкцию с противительностью, но сама по себе безусловным проектором не является, т. к. открывает возможность для такого развертывания, но не является достаточным условием для него.

2.2.3. Еще один тип трудностей при разграничении просодической завершенности и просодической незавершенности перед НО связан с иллокутивно неоднородными конструкциями. Такое происходит, в частности, когда перед НО оказывается правая граница отрезка, передающего чужую речь внутри нарратива. В таких случаях просодия отрезка с чужой речью может варьироваться от полного перевоплощения в цитируемого говорящего (просодически стопроцентно прямая речь) до полного отказа от передачи «чужой» просодии (просодически стопроцентно косвенная речь). При этом лексико-грамматические признаки, характеризующие речь как прямую или как косвенную не обязаны согласовываться с просодическими (см. подробнее Кибрик, Подлесская (ред.) 2009: 288–308). Так, в следующем примере встроенная чужая речь перед НО просодически оформлена как прямая — интонационно имитируется речевой акт угрозы с эмфатически выделенным директивом. Так что никаких просодических симптомов, проецирующих продолжение, нет; фрагмент перед НО следует квалифицировать как просодически завершенный:

(9) Sib_08-f

48.	Они ему \угрожали:
49.	..(0.17) «Только –\сломай-ай Люк!»
50.	...(0.78) Но обходилось без \жертв.

Иначе обстоит дело в примере (10). Формально перед НО — отрезок с прямой речью, директив, речевой акт убеждения, однако просодически картина иная: строки 45–46, начало цитируемого отрезка, действительно имитируют прямую речь, однако строка 48 — парцеллированное прямое дополнение — произносится с редукцией, ускоренным темпом и слабовыраженным ровным движением тона. Здесь говорящая уже начала выходить из режима цитирования и приступила к встраиванию отрезка в локальную структуру нарратива, добавляя этому отрезку просодических симптомов незавершенности:

(10) Sib_12-f

45.	...(0.61) –Она говорила,
46.	..(0.10) «/\Во-от!,
47.	..(0.06) /\посмотри-ите¡!
48.	..(0.30) /Какие –дома.»,
49.	но /мы с \Галечкой не /смотрели ни на \Кремль,
50.	..(0.32) ни на \какие дома,

Такого рода «сходство до степени смешения» просодических стратегий нуждается в дополнительном изучении; во всяком случае, в таких контекстах просодия часто не предоставляет слушающему убедительных сигналов, которые бы однозначно проецировали или, наоборот, блокировали бы дальнейшее развертывание дискурсивной структуры.

2.2.4. Наконец, очевидным образом, группу проблемных случаев пополняют контексты, где в просодическую реализацию НО-конструкций вмешивается речевой сбой:

(11) 04-m_Pr-R

14.	..(0.06) Продавец ему-у сказал очень большую /с-сумму,
15.	...(0.65) н-но-о’ ==
16.	..(0.07) и у= || столько –денег у него не \оказалось.

Таковы, в самом сжатом изложении, ситуации, в которых однозначная характеризация фрагмента, предшествующего НО, как просодически завершенного или просодически незавершенного наталкивается на определенные трудности. Обратимся теперь к случаям, когда на просодический статус фрагмента, предшествующего НО, оказывают влияние факторы, лежащие вне просодии.

3. Лексические, грамматические и дискурсивные факторы, коррелирующие с просодической завершенностью/незавершенностью

Просодические и не просодические проекторы могут совместно вносить вклад в развертывание локальной структуры дискурса.

В частности, некоторые лексико-грамматические явления могут сужать спектр просодических возможностей. Так, наличие в отрезке, предшествующем НО, союзов или иных лексико-грамматических средств, формирующих уступительный или уступительно-ограничительный протазис, способствует просодической незавершенности этого отрезка, ср. конструкцию *не только…но и*:

(12) FS_25-m_Sp

36.	<однако | значит> алкоголь позволяет не /т<u>о</u>лько увидеть красивое в женщине,
37.	но и ….(1.64) собственно говоря …(0.73) иногда помогает этой красоты \избеж<u>а</u>ть.

Однородные определения в препозиции практически не допускают просодической завершенности перед НО — первое определение бывает либо атоническим, либо реализуется с восходящим акцентом:

(13) FS_19-f_Sp

2.	… Это было-о … два месяца /наз<u>а</u>д,
3.	…. в эээ … /стр<u>а</u>нной,
4.	но чудесной стране /Б<u>е</u>льгия,
5.	.. в городе /Льеж,

Заметим, что для постпозитивных однородных определений такого ограничения нет — второй конъюнкт (см. строку 31 в примере ниже) может быть добавлен как парцеллированный после просодически завершенного первого:

(14) 10-f_Pr-T

29.	..(0.23) 'И подарил /жене на –день /рождения /офигительную \ **маш<u>и</u>нку**!
30.	**М-м<u>а</u>ленькую**!
31.	..(0.20) Но /очень \дорог<u>у</u>ю.

Просодическиая завершенность/незавершенность может обуславливаться и семантическими факторами. В частности, просодической незавершенности перед НО способствует так называемое иллокутивное употребление союза (см. Апресян и др. 2010: 224–226). Так, в следующем примере противительное

отношение устанавливается не между двумя пропозициями, а между пропозицией — ‘место отмечается конусом’ и модусной частью — ‘понятия не имею об этом, но выскажу такое мнение’:

(15) FS_18-f_Sp

157.	” поэ= || поэтому \понятия не имею,
158.	что это /такое,
159.	но в общем .. видимо ” м-место —
160.	где нужно /объехать,
161.	— обычно .. отмечается этим \конусом.

Но самой яркой тенденцией, обнаруженной на нашем материале, оказалась зависимость между просодической завершенностью фрагмента перед НО и коммуникативной сложностью фрагмента, следующего за НО. Более точно эту зависимость можно выразить следующим образом. В тех случаях, когда в составе фрагмента, следующего за НО, имеется препозитивная сентенциальная составляющая с тематическим статусом, говорящий стремится оформить фрагмент, предшествующий НО, как просодически завершенный. Или, иначе говоря, если после НО следует фрагмент со сложной дискурсивной структурой, то говорящий стремится оформить этот фрагмент как отдельную иллокуцию. Обычно осложнителями являются препозитивные обстоятельственные клаузы, ср. просодическую завершенность перед НО в (16), строки 2–3, и в (17), строки 14–15:

(16) FS_15-m_Sp

1.	.. Однажды утром я собирался ехать в \университет.
2.	… И-и ехал я на \машине.
3.	.. Но когда я открыл /дверь,
4.	ээ .. то-о /обнаружил,
5.	что-о замок там /примёрз’,
6.	.. и-и дверь не \закрывалась потом.

(17) 01-f_Ski-R

13.	..(0.27) и /\потом ему опять чего-то в голову /ударило,
14.	и он /поехал ка= || опять кататься на \лыжах.
15.	..(0.21) Но так как \о-он \был /пьяный,
16.	…(0.87) /ему…(0.54) ’ ..(0.16) было это сделать \тяжело.

Помимо обстоятельственных клауз, препозитивными осложнителями с тематическим статусом могут быть сентенциальные темы-заголовки, в которых вводится общее описание или ярлык некоторого положения дел, существо дела же раскрывается в последующей порции дискурса. Таковы, например, «катафорические» ярлыки *минус* в (18) и *торжество* в (19):

(18) FS_26-m_Sp

15.	Она очень сильно разогревает .. {СМЕХ} .. как бы /мышцы,
16.	и они как бы начинают более-менее так \работать,
17.	ну и \боль снимает.
18.	... \Во-от.
19.	.. Но у неё есть одно /минус,
20.	она очень \жгучая там,

(19) Sib_12-f

66.	..(0.21) *и мы ничего не \видели*,
67.	потому что *мы никуда не \смотрели*,
68.	мы /только смотрели под \ноги.
69.	Но зато /вечером ..(0.29) у нас было **такое** /торжество,
70.	когда мы эт= ..(0.36) =ти фантики /\разглаживали,
71.	..(0.10) и ..(0.29) \мечтали,
72.	как мы /**поменяемся**,,,

Конструкции с препозитивной сентенциальной темой широко представлены в нашем материале, что позволяет количественно оценить меру их влияния на просодическую завершенность фрагмента перед НО. В корпусе «Рассказы сибиряков о жизни» из 21 вхождения союза НО 4 случая с препозитивной сентенциальной темой (2 — с обстоятельственной клаузой и 2 — с сентенциальной темой-заголовком). Во всех этих случаях фрагмент, предшествующий НО, реализуется как просодически завершенный. В устной части корпуса «Веселые истории из жизни» из 53 вхождений союза НО 10 случаев с препозитивной сентенциальной темой (8 — с обстоятельственной клаузой и 2 — с сентенциальной темой-заголовком). Из этих10 случаев — 9 реализованы с просодической завершенностью перед НО и 1 случай имеет спорную реализацию (см. выше раздел 2.2). В корпусе «Истории о подарках и катании на лыжах» из 32 вхождений союза НО, 9 случаев с препозитивной сентенциальной темой (7 — с обстоятельственной клаузой и 2 — с сентенциальной темой-заголовком). Из этих 9 случаев — 6 реализованы с просодической завершенностью перед НО, 2 — с просодической незавершенностью и 1 случай имеет спорную реализацию. Таким образом, наши корпусные данные убедительно показывают, что конструкции, в которых после НО следует неэлементарный фрагмент с препозитивной сентенциальной темой, устойчиво тяготеют к тому, чтобы фрагмент перед НО артикулировался как просодически завершенный, не проецирующий продолжения.

Обратимся теперь к более общим количественным наблюдениям.

4. Количественные данные и итоговые обобщения

Прежде всего, сравним общую частотность употребления союза НО в наших корпусах и в НКРЯ. В таблице 1 приводятся частоты с приведением к миллиону словоупотреблений (ipm) по данным основного корпуса, устного и мультимедийного с ограничениями на типы текстов[3]:

Таблица 1

Корпус	Число слов в корпусе	НО, всего	НО, ipm	95%-ный доверительный интервал	
				от	до
НКРЯ:					
НКРЯ-основной нехудож.: официально-деловая речь	4 336 610	7 584	1 748,8	1 709,8	1 788,7
НКРЯ-основной нехудож,обиходно-бытовая речь	4 849 097	30 482	6 286,0	6 216,1	6 357,0
НКРЯ-устный, устная публичная речь	6 411 813	40 892	6 377,6	6 316,2	6 439,6
НКРЯ-устный, устная публичная речь, только: доклад | конференция | круглый стол | лекция | семинар	1 051 507	6 840	6 504,9	6 352,6	6 660,9
НКРЯ — МУРКО устная непубличная речь | устная публичная речь	827 625	5 930	7 165,1	6 985,0	7 349,7
Spokencorpora:					
Рассказы сибиряков о жизни	5 000	21	4 200,0	2 669,7	6 533,0
Веселые истории из жизни, устные	10 000	53	5 300,0	4 011,2	6 982,0
Истории о подарках и катании на лыжах	4 500	32	7 111,1	4 949,6	10 151,0
Веселые истории из жизни, письменные	700	31	4 428,6	3 062,7	6 362,7

В целом, можно говорить о том, употребление союза НО больше характерно для устных текстов, чем для письменных, и больше характерно для неформальных типов речевой коммуникации. Наблюдаемые в наших рабочих данных интервалы частот согласуются с интервалами, полученными на существенно бóльших объемах НКРЯ, что позволяет надеяться, что обнаруженные нами тенденции не являются локальным выбросом, а характеризуют устную русскую речь в целом.

---

[3] Выражаю искреннюю признательность А. Ч. Пиперски, который очень помог с корпусной статистикой и чье заинтересованное отношение к сути дела выходит далеко за пределы простой профессиональной солидарности.

Сравним теперь распределение просодически завершенных и просодически незавершенных фрагментов перед НО в трех устных рабочих корпусах — с одной стороны, и распределение НО в начале предложения и НО не в начале предложения в письменном подкорпусе «Веселых историй из жизни» и в основном корпусе НКРЯ — с другой. Запятая и точка являются прототипическими пунктуационными знаками, которые в письменной речи используются перед НО. В «Веселых историях из жизни» других знаков в принципе нет, а в НКРЯ — почти нет. Справедливости ради, отметим, что в исторической перспективе, особенно в авторских редакциях, обнаруживаются редкие экземпляры нестандартной пунктуации. Так в хрестоматийных строфах «Евгения Онегина» перед НО обнаруживаем сочетание знаков «!..», воплощавшее графическую комбинацию восклицательного знака и многоточия[4]. Однако для современного узуса — это, конечно, экзотизм:

(20) Качая важно головою,
Соседи шепчут меж собою:
Пора, пора бы замуж ей!..
Но полно. Надо мне скорей
Развеселить воображенье
Картиной счастливой любви.

В таблице 2 показано — в абсолютных цифрах и в долях от общего числа НО-конструкций — насколько часто фрагменты перед НО оказываются просодически завершенными, просодически незавершенными или не поддаются однозначной интерпретации (такие случаи описаны выше в разделе 2.2):

Таблица 2

Корпус	НО, всего	просодическая завершенность перед НО		просодическая НЕзавершенность перед НО		Спорная просодия	
			доля от общего числа НО в корпусе		доля от общего числа НО в корпусе		доля от общего числа НО в корпусе
Рассказы сибиряков о жизни	21	5	**23.8%**	11	**52.4%**	5	**23.8%**
Веселые истории из жизни, устные	53	21	**39.6%**	22	**41.5%**	10	**18.9%**
Истории о подарках и катании на лыжах	32	10	**31.2%**	11	**34.4%**	11	**34.4%**

В таблице 3 показано — в абсолютных цифрах и в долях от общего числа НО-конструкций — насколько часто НО в письменном тексте появляется в начале

[4] Выражаю искреннюю признательность Н. В. Перцову за консультацию по пушкинской пунктуации.

или не в начале предложения (поиск по НКРЯ осуществлялся через запросы «НО с заглавной буквы», «НО не с заглавной буквы» и, дополнительно, «НО — после запятой»; запрос «НО после многоточия, к сожалению, в НКРЯ недоступен»):

Таблица 3

Корпус	НО, всего	НО с заглавной буквы		НО не с заглавной буквы»	
			доля от общего числа НО в корпусе		доля от общего числа НО в корпусе
НКРЯ-основной нехудож.: официально-деловая речь	7584	1487	**19,6%**	6097	**80,4%**
				в том числе, после запятой:	
				5535	***73,0%***
НКРЯ-основной нехудож,: обиходно-бытовая речь	37454	9328	**24,9%**	28126	**75,1%**
				в том числе, после запятой:	
				24979	***66,7%***
Веселые истории из жизни, письменные	31	8	**25,8%**	23	**74,2%**
				в том числе, после запятой:	
				23	***74,2%***

Как показывают таблицы 2 и 3, доли случаев с НО в начале предложения в письменной речи по НКРЯ и по нашему экспериментальному корпусу находятся в сравнимых интервалах — 20–25% от числа вхождений НО. При этом заметно, что подкорпус официально-деловой речи НКРЯ дает более низкую долю (19,6%), чем подкорпус обиходно деловой речи (24,9%). Если сравнить результаты по письменным и устным версиям «Веселых историй», рассказанных одними и теми же носителями, то тоже заметна разница: доля случаев с точкой перед НО в письменных версиях рассказов составляет 25,8%, тогда как доля случаев просодической завершенности перед НО в устных версиях составляет 39,6%. И в целом, по трем исследованным корпусам, доли случаев просодической завершенности перед НО выше, чем доли случаев с точкой перед НО в письменных версиях.

Таким образом, если обобщить наши количественные наблюдения, можно заключить, что и само употребление союза НО и тенденция к дискурсивной самостоятельности компонента, вводимого НО (выделение его в отдельную илллокуцию) больше характерны для устных текстов, чем для письменных, и больше характерны для неформальных типов речевой коммуникации, чем для формальных. По-видимому, в письменных текстах более влиятельным оказывается прототип бинарной конструкции с НО, где оба компонента объединены в единую иллокуцию, и запятая является стандартным пунктуационным проектором, «обещающим» читающему появление в тексте фрагмента, связанного с текущим. В устной, и особенно, в устной неформальной коммуникации, используются более разнообразные стратегии, комбинирующие разные типы проекторов — лексико-грамматических и просодических.

Литература

1. *Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л., Санников В. З.* (2010) Теоретические проблемы русского синтаксиса: Взаимодействие грамматики и словаря. Москва: Языки славянских культур.
2. *Брызгунова Е. А.* (1982) Интонация, Русская грамматика, том 1, Наука, Мо-ва, сс. 103–118.
3. *Кибрик А. А., Подлесская В. И.* (Ред.) (2009) Рассказы о сновидениях: Корпусное исследование устного русского дискурса. Москва: Языки славянских культур.
4. *Санников В. З.* (1989) Русские сочинительные конструкции. М.: Наука.
5. *Янко Т. Е.* (2008) Интонационные стратегии русской речи в сопоставительном аспекте. Москва: Языки славянских культур.

References

1. *Apresjan, J. D.; Boguslavskij, I. M.; Iomdin, L. L.; Sannikov, V. Z.* (2010) Teoreticheskie problemy russkogo sintaksisa: Vzaimodejstvie grammatiki i slovarja [Theoretical problems of Russian Syntax: Interaction grammar and lexicon]. Moskva: Jazyki Slavjanskix Kul'tur.
2. *Auer, Peter* (2002) Projection in interaction and projection in grammar." // Interaction and linguistic structures # 33.
3. *Bryzgunova E. A.* (1982) Intonation [Intonatsiya], Russian Grammar [Russkaya grammatika]. Vol. 1, Nauka, Moscow, pp. 98–118.
4. *Janko T. E.* (2008) Intonacionnye strategii russkoj rechi v tipologicheskom aspekte [Intonational strategies in spoken Russian from a comparative perspective]. Moskva: Jazyki Slavjanskix Kul'tur.
5. *Kibrik A. A., Podlesskaya V. I. [Eds.]* (2009) Rasskazy o snovidenijax: korpusnoe issledovanie usntogo russkogo diskursa [Night Dream Stories: A corpus study of spoken Russian discourse]. Moscow: Jazyki Slavjanskix Kul'tur.
6. *Malchukov Andrej* (2004) Towards a Semantic Typology of Adversative and Contrast Marking. Journal of Semantics 21, 177–198.
7. *Sannikov, Vladimir* (1989) Russkie sočinitel'nye konstrukcii [Russian coordinate constructions]. Moscow: Nauka.
8. *Spokencorpora* (2013) Prosodically Annotated Corpus of Spoken Russian (PrACSRuss). Pilot version. Online: http://spokencorpora.ru.

ВЕРБАЛЬНАЯ РАБОЧАЯ ПАМЯТЬ И ЛЕКСИКО-ГРАММАТИЧЕСКИЕ СИГНАЛЫ РЕЧЕВЫХ ЗАТРУДНЕНИЙ: ДАННЫЕ РУССКОГО МУЛЬТИМОДАЛЬНОГО КОРПУСА[1]

Потанина Ю. Д. (yulia.potanina.msu@gmail.com)
МГУ имени М. В. Ломоносова, Москва, Россия

Подлесская В. И. (vi_podlesskaya@il-rggu.ru)
РГГУ, РАНХиГС, Москва, Россия

Федорова О. В. (olga.fedorova@msu.ru)
МГУ имени М. В. Ломоносова, Институт языкознания РАН, РАНХиГС, Москва, Россия

Хорошо известно, что объем вербальной рабочей памяти коррелирует с индивидуальными различиями испытуемых при понимании речи (Daneman, Carpenter 1980). В тоже время связь объема рабочей памяти и речепорождения пока изучена относительно слабо. В настоящей работе предпринята попытка восполнить этот пробел и проверить гипотезу о корреляции между объемом рабочей памяти и количеством лексико-грамматических маркеров затруднений при порождении спонтанного пересказа. Девятнадцать русскоязычных испытуемых приняли участие в двух тестах: тесте «Speaking span», в котором им был измерен объем рабочей памяти, и тесте на порождение речи, основанном на пересказе «Фильма о грушах» (Chafe 1980). Собранные пересказы были расшифрованы и вручную аннотированы с точки зрения лексико-грамматических сигналов речевых затруднений. Оказалось, что количество сигналов речевых затруднений в пересказах колеблется от 0,77 до 8,58 на 100 слов, что соответствует средним значениям, характерным для спонтанной речи. Проведенный статистический анализ показал, что объем рабочей памяти действительно коррелирует с беглостью речи, измеренной в числе лексико-грамматических маркеров речевых затруднений.

Ключевые слова: вербальная рабочая память, порождение речи, беглость речи, лексико-грамматические сигналы речевых затруднений

[1] Работа выполнена при финансовой поддержке РНФ (проект № 14-18-03819).

VERBAL WORKING MEMORY AND SPEECH PRODUCTION DIFFICULTIES: DATA FROM RUSSIAN MULTIMODAL CORPUS

Potanina Y. D. (yulia.potanina.msu@gmail.com)
Lomonosov Moscow State University, Moscow, Russia

Podlesskaya V. I. (vi_podlesskaya@il-rggu.ru)
RSUH, RANEPA, Moscow, Russia

Fedorova O. V. (olga.fedorova@msu.ru)
Lomonosov Moscow State University, Institute of Linguistics RAS, RANEPA, Moscow, Russia

It's a well-known fact that working memory capacity correlates with individual differences in comprehending speech (Daneman, Carpenter 1980). At the same time, the relationship between working memory capacity and speech production remains relatively unexplored. In this paper, we attempt to partially fill the gap and check the hypothesis about correlation between working memory capacity and number of lexical and grammatical markers of difficulties in production of spontaneous narratives. 19 Russian participants took part in two tests: the "Speaking Span" test by which we have measured their working memory capacity and the speech production test based on retelling the Pear Film (Chafe 1980). The Speaking Span test was designed in (Daneman and Green 1986) for English-speaking individuals: during the test increasingly longer sets of words are presented to participants; at the end of each set, they are supposed to use each word to generate a separate sentence (the word should be in the same grammatical form as it has been presented). Speaking span is measured as the maximum number of semantically and grammatically correct sentences produced in the experiment. This test was adapted to Russian: words in the set were balanced by syntactic categories, frequencies of individual lexemes and frequencies of grammatical forms. Collected narratives have been transcribed and manually annotated for lexical and grammatical markers of production difficulties. The documented number of lexical and grammatical markers of speech production difficulties varied between 0.77 and 8.58 per 100 words, which matches average rates reported previously in the literature. The study demonstrates the statistically significant correlation between working memory capacity measured by the "Speaking Span" test and verbal fluency measured in number of lexical and grammatical markers of production difficulties.

Key words: verbal working memory, speech production, verbal fluency, lexical and grammatical markers of production difficulties

1. Исследования рабочей памяти в когнитивной психологии и психолингвистике

Начало современного этапа изучения памяти в области ***когнитивной психологии*** связано с именем Г. Эббингауза, который в конце XIX в. разработал первые экспериментальные методики ее измерения. Примерно в то же время У. Джеймс предложил разделить память на первичную и вторичную (James 1890). В начале второй половины XX в. в работах Дж. Миллера (Miller 1956) и Р. Аткинсона & Р. Шиффрина (Atkinson, Shiffrin 1968) был сформулирован многокомпонентный подход к памяти. Согласно наиболее известной трехкомпонентной модели Аткинсона и Шиффрина (1968), сначала некоторая входящая информация попадает в сенсорные регистры (так называемый иконический (в зрительной модальности) и эхоический (в слуховой модальности) виды памяти), затем на 10–25 с. переводится в кратковременное хранилище, после чего попадает в долговременную память. Термин «рабочая память»(РП) был впервые использован в (Miller et al. 1960). Употребление термина «рабочая память» вместо «кратковременная память» подчеркивает функциональную значимость этой системы. Согласно современным представлениям, РП состоит из четырех модулей: (1) центрального исполнителя, (2) фонологической петли, (3) визуально-пространственной матрицы и (4) эпизодического буфера (Baddeley 2000). Центральный исполнитель является ядром системы, отвечающим за координацию работы всех ее подсистем, а три других модуля выполняют вспомогательные функции. Вербальная информация, поступающая из первичного сенсорного хранилища, попадает в фонологическую петлю, которая в свою очередь состоит из пассивного фонологического хранилища и подсистемы, обеспечивающей субвокальное повторение, которое препятствует угасанию следа речевого сигнала в памяти; без поддержки системы субвокального повторения информация в фонологическом хранилище угасает примерно через 1,5 с. Зрительная информация, поступающая из первичного сенсорного хранилища, попадает в визуально-пространственную матрицу, также состоящую из двух подсистем — зрительной и пространственной. Многомодальный эпизодический буфер используется для синтеза информации из фонологической петли и визуально-пространственной матрицы, а также для связи с долговременной памятью.

В области ***психолингвистики*** термин «рабочая память» используется с начала 1980-ых гг. Сначала тест на определение объема РП был разработан в области ***понимания речи*** (Daneman, Carpenter 1980). Данеман и Карпентер предположили, что в процессе интерпретации текста в РП происходят процессы, связанные как с пассивным хранением поступающей информации, так и с ее активной обработкой. Разработанный ими тест “Reading span” учитывал оба этих процесса: испытуемые читали отдельные предложения (обработка) и одновременно удерживали в памяти последние слова ранее прочитанных предложений (хранение). По словам Данеман, «теория кратковременной памяти была заменена теорией РП, а методика измерения кратковременной памяти — методикой измерения РП» (Daneman 1994: 443, *перевод наш*). Используя этот тест, авторы показали, что деление людей на «хороших читателей»,

которые умеют эффективно распределять ресурсы РП между хранением и обработкой поступающей информации, и «плохих читателей», которые делают это хуже, имеет под собой экспериментальные свидетельства. Чуть позже аналогичный тест был создан и для изучения процессов ***порождения речи*** (Daneman, Green 1986; Daneman 1991). Тест «Speaking span» состоял в следующем. Для эксперимента было отобрано 100 слов, которые были распределены в блоки по 2, 3, 4, 5 и 6 слов, в эксперименте было использовано по пять блоков каждого типа. Каждое слово появлялось на экране на 1 с. Испытуемый получал инструкцию читать слова, появляющиеся на экране, и, увидев пустой экран, придумывать с каждым прочитанным словом по одному предложению, причем целевое слово в этом предложении должно было стоять в той же грамматической форме. Например, прочитав слова *shelter, muscles* и *dangers*, англоязычный испытуемый произносил: *Trees provide poor* ***shelter*** *during a thunderstorm*; *Mr. Universe has very big* ***muscles***; *There are* ***dangers*** *associated with every occupation*. Объем РП приравнивался к количеству слов, с которыми испытуемый смог придумать предложения.

Оба теста были адаптированы к русскому материалу, об адаптации теста «Reading span» см. работу (Федорова 2003), об адаптации «Speaking span» — работу (Федорова, Потанина 2014).

2. Рабочая память и лексико-грамматические сигналы речевых затруднений

Еще в работе (Daneman 1991) было показано, что объем РП, измеренный с помощью теста «Speaking span» коррелирует с беглостью речи. В работе (Федорова, Потанина 2014) эта корреляция была проверена на русском материале. Исследование, проведенное с 32 испытуемыми, состояло из пяти тестов: 1) «Speaking span»; 2) порождение речи; 3) чтение вслух; 4) тест с провоцированием испытуемых к оговоркам; 5) чтение скороговорок. В результате были получены значимые корреляции между объемом РП и (1) количеством слов в тесте на порождение речи ($cor = 0{,}522$, $p\text{-value} < 0{,}05$); (2) временем чтения вслух ($cor = -0{,}704$, $p\text{-value} < 0{,}01$); (3) количеством оговорок в тесте на оговорки ($cor = -0{,}706$, $p\text{-value} < 0{,}01$); (4) временем чтения скороговорок ($cor = -0{,}500$ и $-0{,}471$ для первой и второй попыток прочтения, соответственно, $p\text{-value} < 0{,}05$). Однако, в отличие от результатов Данеман, значимой корреляции между объемом РП и количеством ошибок при чтении вслух (как для художественного текста, так и для скороговорок) обнаружить не удалось. Цель данного исследования состоит в изучении взаимосвязи объема РП и маркеров речевых затруднений при порождении устных пересказов. Мы предполагаем, что объем РП будет обратно коррелировать с количеством маркеров речевых затруднений, т. е. при относительно большом объеме РП испытуемого мы получим относительно небольшое количество маркеров речевых затруднений в его речи, и, наоборот, небольшой объем РП будет сопровождаться большим количеством речевых затруднений.

2.1. Сбор подкорпуса «Рассказы о грушах»

Исследование по определению взаимозависимости между объемом РП и маркерами речевых затруднений было выполнено на материале 19 пересказов известного «Фильма о грушах» У. Чейфа (Chafe 1980). Корпус, отобранный для проведения данного исследования, состоял из двух частей. Первую часть (9 пересказов) составили записи, собранные по традиционной процедуре: каждый испытуемый смотрел незнакомый ему шестиминутный фильм, а затем пересказывал его содержание другому человеку, который этот фильм не видел. Вторая часть пересказов (10 записей) была взята из подкорпуса «Рассказы и разговоры о грушах», являющегося частью проекта РНФ «Язык как он есть: русский мультимодальный дискурс»,включающего 24 записи общей длительностью 10 часов и объемом более 110 000 слов. В каждой записи принимали участие 4 человека с заранее распределенными ролями Рассказчика, Комментатора, Пересказчика и Слушателя; для настоящего исследования из этих записей были отобраны части, в которых Рассказчик рассказывал о содержании фильма Пересказчику в присутствии Комментатора.

Все отобранные пересказы представляют собой монологическую речь продолжительностью от 2 мин. 50 с. до 7 мин. 37 с., объем каждого пересказа лежит в интервале от 279 до 967 слов. Суммарная длительность записей составила 84 мин., суммарный объем — 10 700 слов. С каждым испытуемым был проведен тест по определению объема его РП по описанной выше процедуре; объем РП подсчитывался в процентах и находится в интервале от 49 % до 84 %.

2.2. Аннотация лексико-грамматических маркеров речевых затруднений

Во всех собранных пересказах была вручную произведена разметка лексико-грамматических сигналов речевых затруднений. Долексические сигналы, такие как заполненные паузы и фонологически не мотивированные удлинения звуков, нами не рассматривались, возможная их связь с объемом РП должна составить в перспективе предмет отдельного исследования. Аннотирование маркеров речевых затруднений производилось на основе классификации, предложенной в работах (Кибрик, Подлесская (ред.) 2009; Подлесская 2013; Подлесская 2014) с некоторыми уточнениями. Нами учитывались следующие явления[2]:

а) ***Микрокоррекции*** — самоисправления говорящего в пределах ЭДЕ, напр., замена слова в (1):

---

[2] Индекс примера содержит отсылку к ярлыку пересказа в составе выборки. Примеры из текстов выборки даны в дискурсивной транскрипции, разработанной в (Кибрик, Подлесская (ред.) 2009) и принятой также в коллекции корпусов устной речи (Spokencorpora 2013). В соответствии с базовыми принципами этой транскрипции тексты были разбиты на элементарные дискурсивные единицы (ЭДЕ) — минимальные кванты дискурса, обладающие грамматической и коммуникативно-просодической целостностью; строка транскрипта соответствует одной ЭДЕ.

(1) #16
сначала показывается || ээ ээ описывается /пейз<u>а</u>ж,

б) ***Макрокоррекции*** — самоисправления говорящего, затрагивающие более одной ЭДЕ. Напр., в (2) строка 2 отменяется, так как она является преждевременной попыткой, удачно реализованной лишь позднее, в строке 4:

(2) #5
ээ он на неё /→засм<u>а</u>тривается,
/\и-и ээ значит колесо натыкается ==
у него во-п= || сначала сдувает его /шл<u>я</u>пу светл=|| светлую,
а потом он колесом натыкается на-а /к<u>а</u>мень,

в) ***Гибридная коррекция*** — строка, планировавшаяся изначально как единая ЭДЕ, обрывается и достраивается в отдельной ЭДЕ, при этом материал исходной ЭДЕ не отбраковывается полностью. Так, в (3) в строке 2 происходит коррекция сказуемого, но эта строка разделяет со строкой 1 подлежащее и содержит анафорическую отсылку к прямому дополнению *груши*:

(3) #18
/мужчина собирает \груши с де= ==
обрывает их с /д<u>е</u>рева,
складывает в большие /→корз<u>и</u>ны,,,

г) ***Маркеры препаративной подстановки*** — слова, обычно местоименного происхождения, которые выполняют функцию временного замещения искомого выражения при затрудненном поиске. Типичным для дескриптивных фрагментов дискурса, напр., оказывается использование в этой функции местоименного прилагательного *такой*, проецирующего грамматическую форму отложенного атрибута, ср. женский род, родительный падеж в (4), типичным является и дублирование акцента — на маркере и на отложенном атрибуте:

(4) #16
мальчик \такой || \европ<u>е</u>йской /вн<u>е</u>шности,

д) ***Маркеры нечеткой номинации***, в частности, так называемые маркеры-аппроксиматоры (см. (Подлесская 2013)), которые сопровождают «пробные» попытки вербализовать некоторый смысл, сигнализируя о том, что предпринятая попытка нуждается в обобщении или уточнении, но адекватного языкового выражения в арсенале говорящего в данный момент не нашлось. Так, в (5) после маркера препаративной подстановки *такая* в строке 1, следуют два маркера нечеткой номинации — *знаешь* во второй строке и *что ли* — в третьей:

(5) #16
ээ перед передним /колесом у него есть ээ такая вот ==
/знаешь¿
/→седушка что ли…

е) ***Маркеры эмоциональной реакции на речевую проблему*** — формирующие автономную реплику междометия, сигнализирующие об эмоциональном состоянии говорящего в связи с обнаружением сбоя. Это может быть, напр., удивление при обнаружении собственной ошибки, как *ой* в (6), строка 4, или досада от невозможности выбрать адекватную номинацию, как *уфф* в (7), строка 2. В (6) маркер эмоциональной реакции комбинируется с макрокоррекцией (обрыв строки 2), в (7) — с микрокоррекцией в строке 1 и маркером нечеткой номинации, конструкцией «X не X» в строке 3:

(6) #16
на первом /плане,
ээ яблоневое ==
ой \нет!,
\грушёвое \дерево.

(7) #16
у одного б-была-а || ну такое /приспособление,
уфф!,
как теннис не /теннис,
вот теннисная … ээ /ракетка небольшая,
и на ней на верёвочке вот теннисный \мячик.

ж) ***Оффлайн коррекции, или редактирование*** — отложенные эксплицитные самоисправления, не нарушающие ни грамматическую, ни просодическую когерентность текста, постфактум направляющие слушающему инструкцию о том, какого рода ошибка содержалась в предшествующем фрагменте, ср. (8), где редактирование происходит в строках 2–3:

(8) #18
мимо проходит какая-то /→коза-а,,,
ээ точнее проходит мимо \↑мужчина,
с \козой,

Редактирование нередко сочетается с другими сигналами речевых затруднений. Так, в (9) «внахлест» реализуются макрокоррекция и редактирование — в строке 3 эксплицитно редактируется строка 1 и подтверждается правильность смены номинации, произошедшей в строке 2 (правильно *девочка*, а не *девушка*), при этом строка 2 обрывается (макрокоррекция) и реализуется уже как строка 4, после редактирования:

(9) #18
видит как навстречу /–ему едет \девушка на \↑велосипеде,,,
ээ ну девочке где-то-о ==
\девочка даже скорее,
девочке лет ну /-тринадцать-\четырнадцать,

з) ***Сплит*** — разрыв текущей ЭДЕ в связи необходимостью ввести некоторый фрагмент дискурса безотлагательно. Так, в (10) разорвана клауза *он аккуратно её протирает*, с тем, чтобы внутрь её, не дожидаясь завершения, вставить адвербиальное уточнение *бережно*:

(10) #9
он мм аккуратно её —
\бережно,
— /п-протирает,
и кладёт в корзину.

Нередко внутри сплита происходит оффлайн коррекция, см. (11), где разрывается клауза *потом проезжает мальчик на велосипеде* в связи с безотлагательной необходимостью заменить *проезжает* на *приезжает*:

(11) #18
/потом проезжает —
\приезжает точнее,
— /мальчик на \велосипеде.

2.3. Анализ и обсуждение результатов

В результате проведенной аннотации мы получили 529 лексико-грамматических сигналов речевых затруднений, следующим образом распределенных по 19 пересказам (см. табл. 1).

Как видно из табл. 1, количество лексико-грамматических сигналов речевых затруднений в пересказах «Фильма о грушах» колеблется в пределах от 0,77 до 8,58 на 100 слов (среднее 4,57) и от 0,71 до 14,46 в минуту (среднее 6,02), что в целом соответствует средним значениям, характерным для спонтанной речи. Согласно данным, представленным в работе (Подлесская 2014), частотность самоисправлений обычно фиксируется в интервале от 1 до 7 самоисправлений в минуту; в диалогах она обычно выше, чем в монологической речи, а в бытовой речи — выше, чем в официальной. Так, для китайских диалогов приводятся сведения о 5,4 случаях самоисправлений в минуту (Tseng 2006); для английских устных пересказов — 1,9–3,7 на 100 слов (Fraundorf, Watson 2008); для венгерских диалогов — 3,8 в минуту (Németh2012); для японских монологов — 1,2 на 100 слов (Maruyama, Sano 2006); для спонтанных диалогов в иорданском варианте арабского языка — 1,6 на 100 слов (Al-Harahsheh 2015).

Таблица 1. Основные количественные характеристики аннотированных пересказов

#	Объем РП	Кол-во затруднений	Длительность (в секундах)	Объем (в словах)	Кол-во затруднений (на 100 слов)	Кол-во затруднений (в минуту)
1	84	34	322	709	4,80	6,33
2	82	51	385	956	5,33	7,94
3	82	34	336	733	4,64	6,07
4	76	7	177	361	1,94	2,37
5	75	14	174	372	3,76	4,83
6	74	2	170	279	0,77	0,71
7	71	10	345	584	1,71	1,74
8	70	11	249	351	3,13	2,65
9	70	28	216	481	5,82	7,78
10	64	42	221	745	5,64	11,41
11	64	12	214	419	2,86	3,36
12	63	13	205	347	3,75	3,80
13	62	17	183	294	5,78	5,57
14	59	72	422	967	7,45	10,24
15	58	18	224	333	5,41	4,83
16	56	24	221	449	5,36	6,52
17	52	40	457	963	4,16	5,26
18	49	60	249	699	8,58	14,46
19	48	40	285	664	6,02	8,42

Проведя статистический анализ, мы нашли значимую корреляцию между объемом РП и отношением числа сигналов речевых затруднений к объему пересказа (в словах): cor = −0,483*, p-value< 0,05, см. рис. 1.

Таким образом, мы подтвердили гипотезу о том, что вербальный объем РП является фактором, коррелирующим с количеством лексико-грамматических сигналов речевых затруднений, т. е. с плавностью речепорождения: чем больше у испытуемого объем РП, тем меньше подобных маркеров обнаруживается в его спонтанной монологической речи.

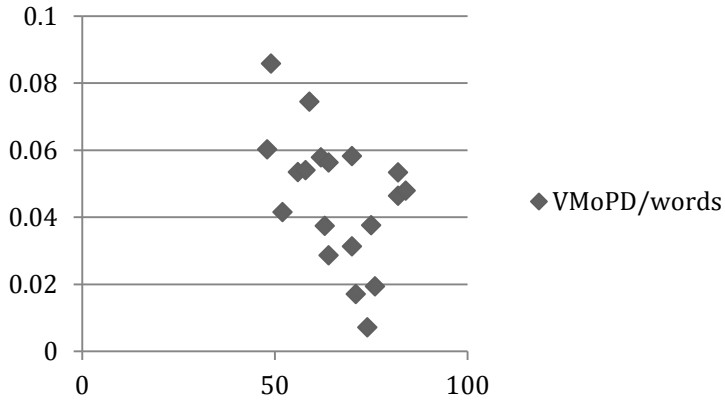


Рис. 1. Точечная диаграмма для объема РП (ось X) и отношения числа сигналов речевых затруднений к объему рассказа (ось Y)

2.4. Перспективы дальнейших исследований

В настоящей работе мы показали, что объем РП коррелирует с индивидуальными различиями при речепорождении. Более конкретно, впервые на материале русского языка была продемонстрирована значимая корреляция между вербальным объемом РП и числом лексико-грамматических сигналов речевых затруднений при порождении пересказа. Данное исследование является пилотным в области анализа взаимосвязи объема РП и маркеров вербальных затруднений, оно проведено на сравнительно небольшом объеме материала и оставляет множество вопросов для дальнейшего изучения. Опишем некоторые из них более подробно.

Во-первых, нуждается в проверке гипотеза о наличии корреляции между объемом РП, полученным в тесте «Speaking span», и количеством речевых затруднений, которые возникали у испытуемых в ходе прохождения самого этого теста. Процедура прохождения испытуемыми теста «Speaking span» была записана на диктофон, так что расшифровка и анализ лексико-грамматических сигналов речевых затруднений даст ответ на вопрос о наличии корреляций как между объемом РП и количеством речевых затруднений при прохождении теста, так и между речевой продукцией испытуемых в ходе прохождения теста и в ходе пересказа.

Во-вторых, в настоящей работе была рассмотрена взаимосвязь между объемом РП и лексико-грамматическими сигналами речевых затруднений. Как было упомянуто выше, другой класс речевых затруднений включает различные **долексические маркеры**, в первую очередь заполненные паузы и фонологически не мотивированные удлинения звуков. Подобные маркеры более частотны в спонтанной речи, но их связь с объемом РП не так очевидна, так как интуитивно они менее когнитивно трудозатратны. Данная гипотеза также может быть проверена на рассмотренном материале 19 пересказов.

В-третьих, представляет интерес вопрос о взаимозависимости объема РП и количества сигналов вербальных затруднений в ситуации дефицита времени, т. е. в таких тестах, в которых испытуемым нужно уместить как можно больше информации в ограниченный по длительности пересказ. Эти и многие другие вопросы требуют как дополнительных исследований вербальной РП, так и более детального анализа лексико-грамматических сигналов речевых затруднений.

Наконец, обнаружение корреляции между объемом РП и вербальными маркерами речевых затруднений может означать, что (а) низкий объём рабочей памяти увеличивает затруднения при порождении, (б) затруднения снижают объем РП, (в) другой фактор или факторы влияют на рабочую память и речевые затруднения[3]. Чтобы определить, каким образом объем РП связан с речевыми затруднениями, необходимо более детальное исследование механизма синтеза речи и роли рабочей памяти в этом процессе. Кроме того, необходимо более тщательное изучение различных классов речевых затруднений и анализ вклада каждого класса в корреляцию с объемом РП. На наших данных не удалось продемонстрировать значимых корреляций между объемом РП и определенными типами затруднений. Мы предполагаем, что для данного исследования понадобится больший объем корпуса рассказов, в котором будут представлены все известные типы затруднений.

Литература

1. *Al-Harahsheh A. M. A.* (2015), A Conversation Analysis of self-initiated repair structures in Jordanian Spoken Arabic. Discourse Studies 17(4), 397–414
2. *Atkinson R. C., Shiffrin R. M.* (1968), Human memory: A proposed system and its control processes, K. W. Spence, J. T. Spence (eds.) The psychology of learning and motivation: Advances in research and theory, New York.
3. *Baddeley A. D.* (2000), The episodic buffer: A new component of working memory? Trends in Cognitive Sciences 4(11), 417–23.
4. *Chafe W.* (ed.) (1980), The pear stories: Cognitive, cultural, and linguistic aspects of narrative production, Norwood: Ablex.
5. *Daneman M.* (1994), Working memory and language, Language and Speech, 37.
6. *Daneman M.* (1991), Working memory as a predictor of verbal fluency. Journal of Psycholinguistic Research, 20(6), 445–464.
7. *Daneman M., Carpenter P. A.* (1980), Individual differences in working memory and reading. Journal of verbal learning and verbal behavior, 19(4), 450–466.
8. *Daneman M., Green I.* (1986), Individual differences in comprehending and producing words in context. Journal of memory and language, 25(1), 1–18.

[3] Мы выражаем искреннюю признательность анонимному рецензенту, справедливо указавшему на то, что сама по себе корреляция между объемом РП и частотой речевых сбоев не объясняет механизма связи между этими феноменами и нуждается в интерпретации в рамках той или иной модели речепорождения. Эта работа нам еще предстоит.

9. *Fedorova O. V.* (2003), Survey of the State of the Art in Verbal Span Test [Test po opredeleniju ob"ëma operativnoj pamjati: Istorija i sovremennoe sostojanie], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2003" [Komp'iuternaia Lingvistika I Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2003"], Moscow.
10. *Fedorova O. V., Potanina Ju. D.* (2014), Working memory and Russian language: from comprehension to production [Rabochaya pamyat' I russkiy yazyik: ot recheponimaniya k recheporozhdeniyu], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2014" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2014"], Bekasovo.
11. *Fraundorf S. H., Watson D. G.* (2008), Dimensions of variation in disfluency production in discourse, in J.
12. *James W.* (1890). The principles of psychology (Vol. 1). New York: Holt.
13. *Kibrik A. A., Podlesskaya V. I.* [Eds.] (2009), Night Dream Stories: A corpus study of spoken Russian discourse [Rasskazy o snovidenijax: korpusnoe issledovanie usntogo russkogo diskursa]. Moscow: Jazyki Slavjanskix Kul'tur.
14. *Maruyama T. and Sano Sh.* (2006), Classification and Annotation of Self-Repairs in Japanese Spontaneous Monologues, in LPSS — Linguistic Patterns in Spontaneous Speech, Taipei, 283–298.
15. *Miller G. A.* (1956), The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological review, 63(2), 81.
16. *Miller G. A., Galanter E., Pribram K. H.* (1960), Plans and the structure of behavior, New York.
17. *Németh Z.* (2012), Recycling and replacement self-repairs in spontaneous Hungarian conversations, in Proceedings of the First Central European Conference in Linguistics for postgraduate Students, 211–224.
18. *Podlesskaya V. I.* (2014), They shot him dead, oh, no, they knifed him dead with a saber: self-repairs in oral stories [To est', ne ubili, a zarezali sablej: samoispravlenija govorjashhego v ustnyh rasskazah]. Computational Linguistics and Intellectual Technologies: papers from the Annual International Conference "Dialog". V. 13 (20). Moscow: RSUH, 2014, 526–540.
19. *Podlesskaya V. I.* (2013), Vague reference in Russian: evidence from spoken corpora [Nechetkaja nominacija v russkoj razgovornoj rechi: opyt korpusnogo issledovanija]. Computational Linguistics and Intellectual Technologies: papers from the Annual International Conference "Dialog". V. 12 (19). Moscow: RSUH, 2013, 561–573.
20. *Spokencorpora* (2013), Night dream stories and other spoken corpora [Rasskazy o snovidenijah i drugie korpusa zvuchashhej rechi]. Prosodically Annotated Corpus of Spoken Russian. Pilot version. Online: http://spokencorpora.ru
21. *Tseng S.-C.* (2006), Repairs in Mandarin conversation, Journal of Chinese Linguistics 34(1), 80–120.

ВЫЯВЛЕНИЕ МАШИННО-ПЕРЕВЕДЁННЫХ ТЕКСТОВ В КОЛЛЕКЦИИ НАУЧНЫХ ДОКУМЕНТОВ НА РУССКОМ ЯЗЫКЕ

Романов А. В. (romanov@ap-team.ru),
Кузнецова М. В. (kuznetsova@ap-team.ru),
Бахтеев О. Ю. (bahteev@ap-team.ru),
Хританков А. С. (khritankov@ap-team.ru)

Antiplagiat.Research, Москва, Россия;
Московский физико-технический институт
(государственный университет), Москва, Россия

Ключевые слова: обработка естественного языка, статистический машинный перевод, выявление машинного перевода, статистические языковые модели, word2vec

MACHINE-TRANSLATED TEXT DETECTION IN A COLLECTION OF RUSSIAN SCIENTIFIC PAPERS

Romanov A. V. (romanov@ap-team.ru),
Kuznetsova M. V. (kuznetsova@ap-team.ru),
Bakhteev O. Yu. (bahteev@ap-team.ru),
Khritankov A. S. (khritankov@ap-team.ru)

Antiplagiat.Research, Moscow, Russia; Moscow Institute
of Physics and Technology (MIPT), Moscow, Russia

In this paper, we propose a method of machine-translated text detection. By 'machine-translated' texts, we mean, principally, the output of statistical machine translation systems. We focus on syntactic correctness and semantic consistency of sentences that constitute a text. More specifically, we make an attempt of detecting a certain phenomenon typically occurring in machine-translated documents. This phenomenon comprises the cases when small parts of the sentence, correctly translated, are combined together in an improper way. The proposed method is based on a supervised approach with a number of handcrafted features. First, we construct N-gram language models on a set of authentic scientific papers and on a set of machine-generated texts and assess the probability of each sentence according to these models. In addition, we propose N-gram language

models on part-of-speech tag sequences corresponding to the texts given. Furthermore, we explore the effectiveness of features obtained from two trained word2vec (CBOW and skip-gram) models. We assess quality of the method on a sample of Russian scientific papers, and English scientific documents machine-translated into Russian. Preliminary results demonstrate feasibility of the approach.

Key words: natural language processing, statistical machine translation, machine translation detection, statistical language models, word2vec

1. Introduction

Recent advances in the field of statistical machine translation (SMT) and online translation services built upon it have enabled a massive increase in the volume of machine-translated texts available on the Web. Such a technology can be useful for providing information in languages with 'low density' on the Web. For example, automatically translated software documentation [15] becomes more readily available to speakers of these languages. On the other hand, prompt availability of translation systems leads to their potential misuse. Our experience with analysis of student home assignments, course projects and research papers shows that these often contain translated text fragments. While academic papers may contain text either translated by a human or translated with a translation system and post-processed by a human afterwards, student works frequently include fragments translated with one of public SMT systems and left "as is". Such works often lack detailed analysis by a human expert due to large amounts of them or due to the expert's negligence. Moreover, only a portion of text may be machine-translated, thus the problem of detection descends to the sentence level.

Development of automatic methods of machine-translated text detection may help to discover such misuse and is, therefore, of great importance for today's academic community.

Our approach is to develop an algorithm that would be able to classify sentences into two classes: human-written and machine-translated.

In this paper, we focus on detecting *word salad* and *phrase salad* phenomena, which are often produced by machine translation systems. The word salad is commonly defined as a "confused or unintelligible mixture of seemingly random words and phrases; the words may or may not be grammatically correct, but are semantically confused to the point" [22]. The phrase salad denotes text segments in which small parts of sentences are semantically consistent, grammatically and syntactically correct, but are combined together in an improper way, so that the entire sentences become incorrect or absurd.

We propose features obtained from statistical language models (LM) such as N-gram LMs. We construct such LMs on words of the sentences and on sequences of part-of-speech (POS) tags corresponding to them. In addition, we investigate the applicability of the features obtained from trained word2vec (skip-gram and CBOW) models. We conduct a series of experiments on human-written Russian scientific papers, and English papers machine-translated into Russian in order to evaluate the quality of our method.

2. Related work

The problem of translated text detection has already been an active research topic for several years. Early works on this topic [6], [9] aimed to develop a method of automatic evaluation of machine translation systems (in order to track improvements of the system over time or to compare machine-translated texts with reference translations performed by human experts). The authors develop classifiers for distinguishing human-translated sentences from machine-translated ones and propose several groups of features, including language model likelihood scores for sentences, density of function words, features that capture syntactic relationships between words, and others. The topic of automatic SMT quality assessment is further developed in [4], where the authors focus on translation output ranking on a basis of the estimates of sentence grammaticality. In [1] the authors detect output of several SMT systems and conclude the dependence of detection quality and machine translation quality. The majority of works in this area rely heavily on human 'gold-standard' translation availability for source sentences, which is not the case in our task setting.

The development of corpus linguistics has drawn attention to automatic preparation of bilingual corpora. Several methods involve automatic detection of machine-translated text for eliminating low-quality content from parallel corpora. In [2] the authors try to detect the output of phrase-based SMT systems by exploiting their limited potential of word reordering within a sentence. Another approach [20] determines the likelihood of bilingual sentence pairs to be machine-translated using such features as character length ratio, token length ratio etc.

An area related to machine-translated text detection is detection of *translationese*, i.e. detection of overly literal translation performed without taking language features and idioms into consideration, in a collection of authentic documents written by native speakers. The proposed methods [5], [13], [21] are, in principle, related to aforementioned approaches in terms of variety of features used, but the problem setting is different.

Another related field is automatic web spam detection. The problem is similar to the one being considered, as it handles machine-generated texts lacking grammatical correctness. The authors of [11] analyze bigram co-occurrences in order to detect word salad. In [17] and [18] the authors focus on detection of texts generated by specific models (Markov chain generators etc.) These works handle less complicated models of text generation, while SMT systems produce more diversified documents.

Our approach builds upon and extends the method presented in [3], which was designed for English-Japanese pair of languages. In addition to POS tag sequences and LMs, our approach also takes into account some specific characteristics of Russian words (e.g. the case of a noun, the tense of a verb etc.), which enable detection of disagreement in machine-translated sequences.

Moreover, we use features calculated with word2vec CBOW and skip-gram models [14] to capture statistical irregularities of machine-translated texts compared to authentic texts.

3. Method description

3.1. Formal problem statement

Our method is aimed at detection of machine-translated sentences in a mixed sample of authentic human-written sentences in Russian and sentences originally written in another language and machine-translated into Russian. In other words, our task is to determine whether each sentence of a document is machine-translated or not.

Let $D = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^{m}$ be a labeled set of pairs *(object, answer)*, where $\boldsymbol{x_i} = f(x_i)$ is a vector representation of sentence x_i in a feature space $\mathbb{R}^n$ $x_i = \left(x_i^1, \ldots, x_i^{l_i}\right)$; is an ordered sequence of sentence words; $y_i \in Y = \{0, 1\}$ is a class label, where 0 corresponds to the class of authentic sentences and 1 corresponds to the class of machine-translated sentences.

Let $D = D_L \sqcup D_T$ be a partition into a training set D_L and a test set D_T. We set up a problem of finding a classification model $g\colon \mathbb{R}^n \to Y$ that minimizes the empirical loss, which is the aggregated value of loss function $S\colon Y \times Y \to \mathbb{R}$, $S(y_1, y_2) = [y_1 \neq y_2]$, over test set D_T:

$$\hat{g} = \arg\min_{\mathrm{g}} S(g(x_i), y_i \,|D_T) = \arg\min_{\mathrm{g}} \frac{1}{|D_T|} \sum_{i=1}^{|D_T|} [g(x_i) \neq y_i]$$

3.2. Detection word and phrase salad phenomenon

SMT systems are known to generate texts that may lack semantic consistency and grammatical correctness. Basically, this effect takes place for several reasons. Modern SMT systems face many challenges, including homonymy and polysemy disambiguation, sentence structure transformation for languages that are not closely related, and others. Most of these systems handle segments of a sentence in a source language in order to produce a number of phrases in a target language, which are to be combined in a resulting sentence afterwards. Errors in the latter stage lead to a phenomenon known as *phrase salad*. This phenomenon comprises the cases when grammatically correct phrases are combined together in an improper way. Several examples of phrase salad sentences in Russian produced by machine translation are:

(1) *Все это имеет прямое, как а также косвенное влияние на экономическую деятельность и производственных мощностей.*
На практике не существует широкое признание процесс выбора параметра геометрическое распределение.
Масштаб и положение можно принимать любые значения, совместимых с областью временного ряда.

This phenomenon is a natural extension of the concept of *word salad*, which describes the cases when randomly chosen words or words from different domains

constitute a sentence, which becomes a nonsensical set of words. This may frequently be attributed to the output of text generators that use Markov chains or context-free-grammars in their work.

The latter case is easier to detect using statistical language models since N-gram models trained on a set of authentic documents are sensitive to "unlikely" sequences of words. Phrase salad, however, is more subtle as these language models are not able to detect nonsensical phrases of length greater than N.

Based on these observations, we propose features to capture both word salad and phrase salad by estimating the likelihood of sentences according to previously trained statistical language models. These features capture grammatical correctness of the sentence and its lexical integrity. We attempt to build separate models for authentic human-generated text and for machine-translated text since each class of texts may contain particular hidden properties. By contrasting these two sets of features, we can effectively determine whether the sentence is machine-translated or not.

3.3. Lexical features

We build N-gram models for authentic texts and for machine-translated texts in order to compute likelihood of a sentence according to these models:

$$\widehat{p_{LM}}(x) = \widehat{p_{LM}}(x^1, \dots, x^l) = \prod_{i=1}^{l} \widehat{p_{LM}}(x^i | x^{i-1}, \dots, x^{i-N+1}),$$

where N-gram probabilities are estimated from text corpora. We also propose a small positive probability constant estimate for the N-grams that do not occur in the training corpus. A score of the sentence is log likelihood normalized by the length of the sentence.

We train 2- and 3-gram models on two different sets and thus get four different features from them. The length of N-gram is chosen not to be greater than 3 for several reasons. First, we restrict size of our language models for convenience and performance. Second, we take into account the findings of [3], where the authors confirm a negligible effect of extending models to 4-gram and higher.

3.4. Part-of-speech features

Part-of-speech features are also derived from statistical analysis of language models. Computation of this set of features is similar to that of features described in the previous section, but instead of word sequences we use POS tag sequences. Therefore, probability of a POS N-gram is given by:

$$\widehat{p_{POS}}(x) = \widehat{p_{LM}}\left(h(x^1), \dots, h(x^l)\right) = \prod_{i=1}^{l} \widehat{p_{LM}}(h(x^i) | h(x^{i-1}), \dots, h(x^{i-N+1})),$$

where $h: W \to H$ is a part-of-speech tagging function.

We use the following word tags:

- part of speech for all words;
- gender, case and number for nouns;
- gender, case and number for adjectives and participles;
- grammatical person, case and number for personal pronouns;
- grammatical person (or an indicator of infinitive form) for verbs.

We collect this information to discover syntactical disagreement among sentence words, for example, between a noun and a dependent adjective. This approach, as compared to full syntactic parsing, is a trade-off between accuracy of detection and computation time. Computational complexity of POS tagging is linear to the length of the input while polynomial for full syntactic parsing [8], [23]. For instance, when we see an adjective and an adjacent noun in a Russian sentence, this may not be the case of a noun phrase as they may refer to two different adjacent noun phrases. We make an assumption here that this case is less frequent in real sentences, than a potential disagreement between words in machine-translated sentences. We expect this set of features to be helpful in capturing phrase salad phenomenon.

We encode the entire portion of information obtained from POS tagger in a single token, which is used afterwards in construction of a language model. Like in the case of LM lexical features, we use 2- and 3-gram LMs.

3.5. Word2vec features

We train two different word2vec (skip-gram and CBOW) models on a larger separate sample of authentic Russian sentences. According to these models, sentence scores are computed with respect to the log-likelihood of the word appearing in a certain context, and vice versa. Thus, two more features are added to the feature set.

3.6. Classification

The proposed method associates each Russian sentence x with a feature vector $\boldsymbol{x} = f(x) \in \mathbb{R}^{10}$, We apply a random forest classifier to the two-class classification task in the feature space. The classifier choice is justified by the following. First, we have a limited amount of features, which can be processed by tree composition algorithms with high accuracy. Second, it showed the best performance among several classifiers (including linear models and boosting methods) during our experiments.

4. Experiments

4.1. Data preparation

We prepare a collection of human-generated and machine-translated sentences extracted from scientific papers in a specific domain. This corpus is later used for both training of statistical language models and classifier training.

We attempt to confine sentences from both classes to use similar vocabulary for the sake of purity of the experiment.

As a source of human-written texts, we adopt a sample of jurisprudential and sociological papers available online [7] with open access. We obtain machine-translated sentences from a set of English papers of Munich Personal RePEc Archive [16], and translate them into Russian with an online translation service [10]. We focus on a single translation system since the majority of modern SMT systems produce similar output for the Russian language as a target language. We do not consider rule-based translation systems in this task, as they are likely to generate texts with different language phenomena. We also use a sample of 1M articles from the Russian Wikipedia for word2vec model training.

In human-written texts, we select only sentences that do not contain words in English or other non-Cyrillic words. Such words may or may not refer to machine translation errors and can bias our language models. As a result, we prepared the dataset of *300K* authentic sentences and *300K* machine-translated sentences. We used 2/3 of this collection for training of language models and the remaining part for classifier evaluation.

4.2. Experiment setting

We lowercase all the sentences and split them into tokens. We also remove punctuation and replace all numbers with the single token. We discard sentences that contain fewer than 4 tokens because short machine-translated sentences are known to be barely distinguishable from human-generated ones [3].

We use Python pymorphy2 package [12] to retrieve POS tags and relevant information and scikit-learn [19] RandomForestClassifier implementation. We use 5-fold cross-validation technique to train the classifier and to assess quality of the method on the prepared sample. We tune appropriate parameters of the classifier with grid search.

We evaluate the method with F1-measure, as this metrics aggregates the overall quality of two-class classification. Taking into account ability of the random forest classifier to predict probability scores, we also calculate AUC ROC metrics to obtain more stable performance characteristics of our algorithm.

4.3. Experiment results

4.3.1. Overall performance

We conduct a series of experiments with various subsets of the proposed features. Table 1 shows the results of this study. We consider the case when the features are restricted to LM lexical features as the baseline, since a method using these features has been proposed earlier by various authors. The results show the feasibility of our approach and the effectiveness of use of multiple LM-based features.

Table 1. Performance of different feature combinations

Features	F1	AUC ROC
LM lexical (baseline)	0.754	0.816
LM POS	0.727	0.804
word2vec	0.643	0.673
LM lexical + LM POS	0.826	0.907
LM lexical + LM POS + word2vec	**0.832**	**0.912**

4.3.2. Feature importance

We use feature importance estimates of the random forest classifier to understand whether all of the features are useful in the classification task or not. The numerical values are provided in the Table 2. They suggest that the proposed features of all types contribute to the decision made by the classifier and, therefore, are feasible for our task. Moreover, these findings confirm the benefits from constructing models not only on authentic texts, but also on machine-translated texts.

Table 2. Feature importance ratio

Feature	Importance ratio, %
LM lexical 2-gram score on machine-translated texts	26.8
LM POS 3-gram score on authentic texts	13.4
LM POS 3-gram score on machine-translated texts	11.5
LM POS 2-gram score on machine-translated texts	8.0
LM POS 2-gram score on authentic texts	7.5
CBOW score	7.5
LM lexical 3-gram score on machine-translated texts	6.9
Skip-gram score	6.4
LM lexical 2-gram score on authentic texts	6.2
LM lexical 3-gram score on authentic texts	5.9

4.3.3. Error analysis

We make a random subsample of 50 false positive errors and 50 false negative errors of classification in order to analyze the characteristics of misclassified sentences.

Example sentences of both classes are:

(2) *Сопоставление с результатами натурного эксперимента.* (false positive)
При всей своей строгости и лаконичности эта модель обладает существенным недостатком — она является существенно эксплицитной. (false positive)
Так, мысль Раскольникова, что, убив ростовщицу, он уничтожает только «вошь», паразита и, таким образом, совершает не столько преступление, сколько благодеяние, опровергается рядом обстоятельств. (false positive)

В среднем работал по 10 часов в день и 20 процентов работали по 12 часов в день. (false negative)
Этот курс был запущен полностью практически между иностранными группами по 4–6 человек. (false negative)
Преимущества PTC, в частности, темпы и уровни рынка они приводят. (false negative)

Overall, the most common causes of false positive classifications are:
- use of words and word combinations that do not occur in the training corpus (*натурного эксперимента, эксплицитной*);
- use of out-of-domain constructions and personal names (*Раскольникова, «вошь»*).

The most common causes of false negative classifications are:
- low rate of grammatical errors in translated texts;
- low rate of phrase salad.

These findings suggest that the quality of the method may be improved if a larger corpus of higher quality is used for language model training.

5. Conclusion

We propose a method of machine-translated text detection in a collection of scientific papers. The method is based on the supervised approach and operates individual sentences. We train statistical language models and use likelihood estimates of sentences as classification features. Preliminary experiments show feasibility of LM lexical and LM POS features and achieve decent results on a dataset of documents from a specific domain.

As the future work, we plan to improve the quality of our approach by tuning the parameters of the language models we construct. The experiment shows that different language models could catch specific language features that may occur in the output of SMT systems. Therefore, accurate tuning of the parameters is one of the appropriate tasks for future research. Another challenging problem is applicability of our approach to the more general task of machine-generated text detection, especially to detection of the output of context-free-grammar (CFG) based text generators.

References

1. *Aharoni R., Koppel M., Goldberg Y.* (2014), Automatic Detection of Machine Translated Text and Translation Quality Estimation, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, pp. 289–295.
2. *Antonova A., Misyurev A.* (2011), Building a Web-based parallel corpus and filtering out machine-translated text, Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web, Portland, pp. 136–144.
3. *Arase Y., Zhou M.* (2013), Machine Translation Detection from Monolingual Web-Text, ACL (1), Sofia, pp. 1597–1607.
4. *Avramidis E., Popovic M., Vilar D., Burchardt A.* (2011), Evaluate with Confidence Estimation: Machine ranking of translation outputs using grammatical features, Proceedings of the Sixth Workshop on Statistical Machine Translation, Edinburgh, pp. 65–70.
5. *Baroni M., Bernardini S.* (2006), A new approach to the study of translationese: Machine-learning the difference between original and translated text, Literary and Linguistic Computing, Vol. 21(3), pp. 259–274.
6. *Corston-Oliver S., Gamon M., Brockett C.* (2001), A machine learning approach to the automatic evaluation of machine translation, Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, Toulouse, pp. 148–155.
7. *Cyberleninka.ru,* available at: http://cyberleninka.ru/
8. *Earley J.* (1970), An efficient context-free parsing algorithm, Communications of the ACM, Vol. 13(2), pp. 94–102.
9. *Gamon M., Aue A., Smets M.* (2005), Sentence-level MT evaluation without reference translations: Beyond language modeling, Proceedings of EAMT, Budapest, pp. 103–111.
10. *Google Translate,* available at: http://translate.google.com/
11. *Grechnikov E. A., Gusev G. G., Kustarev A. A., Raigorodsky A. M.* (2009), Detection of Artificial Texts [Poisk neestestvennykh tekstov], Proc. 11th All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies" [Trudy XI Vserossiyskoy nauchnoy konferentsii "Elektronnye biblioteki: perspektivnye metody I tekhnologii, elektronnye kollektsii"], Petrozavodsk, pp. 306–308.
12. *Korobov M.* (2015), Morphological Analyzer and Generator for Russian and Ukrainian Languages, Analysis of Images, Social Networks and Texts, Yekaterinburg, pp. 320–332.
13. *Kurokawa D., Goutte C., Isabelle P.* (2009), Automatic detection of translated text and its impact on machine translation, Proceedings. MT Summit XII, The twelfth Machine Translation Summit International Association for Machine Translation hosted by the Association for Machine Translation in the Americas, Ottawa.
14. *Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J.* (2013), Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems, pp. 3111–3119.
15. *MSDN*, available at: http://msdn.microsoft.com/
16. *Munich Personal RePEc Archive,* available at: http://mpra.repec.org/

17. *Pavlov A. S., Dobrov B. V.* (2009), Detecting Web Spam Created with Markov Chains Text Generators [Metod obnaruzheniya poiskovogo spama, porozhdennogo s pomoschyu tsepey Markova], Proc. 11th All-Russian Scientific Conference “Digital Libraries: Advanced Methods and Technologies” [Trudy XI Vserossiyskoy nauchnoy konferentsii “Elektronnye biblioteki: perspektivnye metody I tekhnologii, elektronnye kollektsii”], Petrozavodsk, pp. 311–317.
18. *Pavlov A. S., Dobrov B. V.* (2011), Detecting Mass-Generated Unnatural Texts through Topical Diversity Analysis [Metody obnaruzheniya massovo porozhdennykh neestestvennykh tekstov na osnove analiza raznoobraziya tematicheskoy struktury tekstov], Proc. 13th All-Russian Scientific Conference “Digital Libraries: Advanced Methods and Technologies” [Trudy XIII Vserossiyskoy nauchnoy konferentsii “Elektronnye biblioteki: perspektivnye metody I tekhnologii, elektronnye kollektsii”], Voronezh, pp. 210–218.
19. *Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., ..., Vanderplas J.* (2011), Scikit-learn: Machine learning in Python, The Journal of Machine Learning Research, Vol. 12, pp. 2825–2830.
20. *Rarrick S., Quirk C., Lewis W.* (2011), MT detection in web-scraped parallel corpora, Proceedings of the Machine Translation Summit (MT Summit XIII), Xiamen.
21. *Twitto-Shmuel N., Ordan N., Wintner S.* (2015), Statistical Machine Translation with Automatic Identification of Translationese, EMNLP 2015, Lisbon, p. 47.
22. *Word salad*, available at: http://en.wikipedia.org/wiki/Word_salad
23. *Younger D. H.* (1967), Recognition and parsing of context-free languages in time n 3, Information and control, Vol. 10(2), pp. 189–208.

АВТОМАТИЧЕСКАЯ МОРФОРАЗМЕТКА КОРПУСОВ РУССКОЯЗЫЧНЫХ СОЦИАЛЬНЫХ МЕДИА: ОБУЧЕНИЕ И ОЦЕНКА КАЧЕСТВА

Селегей Д. (danila-slg@yandex.ru)[1],
Шаврина Т. (rybolos@gmail.com)[1],
Селегей В. (Vladimir_S@abbyy.com)[2,3],
Шаров С. (s.sharoff@leeds.ac.uk)[4]

[1]Московский Государственный Университет, Россия
[2]Российский государственный гуманитарный университет, Россия
[3]ABBYY, Россия
[4]Университет Лидса, Великобритания

В статье описывается новый комплекс базовых средств для морфоразметки текстов русскоязычных social media, разработанный в рамках проекта создания мегакорпуса русскоязычного интернета ГИКРЯ. Этот комплекс включает новый tagset, полученный некоторым расширением и адаптаций тагсета, предложенного в Шаров et al., и тестового корпуса (золотого стандарта) современных social media с такой разметкой объемом около 2 млн токенов.

Тагсет, созданный с учетом самых популярных текущих тагсетов РЯ (MULTEXT-East, NLC, разметкой mystem) и сохраняющий совместимость с этими форматами, будет использован для морфоразметки корпуса ГИКРЯ.

Особенностью подхода к применению нового стандарта является полностью автоматический способ разметки тестовых и обучающих корпусов: для этой цели мы использовали результаты синтаксического разбора текстов разных разделов social media с помощью парсера Compreno, используемого по лицензии Compreno Based Research (CBR), с последующим уточнением разметки за счет автоматических процедур коррекции систематических ошибок, возникающих при анализе текстов этого сегмента. Показано, что имеющиеся парсеры (в частности, tnt) показывают при обучении на таком корпусе лучшие результаты для этого сегмента, чем при обучении на других ранее имевшихся в наличии размеченных корпусах (прежде всего, т. н. «снятника» НКРЯ).

Мы полагаем, что одним из наиболее полезных результатов данной работы является появившаяся возможность объективного сравнения результатов работы POS-тегеров на сегменте social media с помощью нового тестового корпуса, который размещен на сайте ГИКРЯ.

Ключевые слова: автоматическая морфоразметка, морфологическая разметка, морфотэггеры для русского языка, язык социальных медиа, обучение морфопарсера

AUTOMATIC MORPHOLOGICAL TAGGING OF RUSSIAN SOCIAL MEDIA CORPORA: TRAINING AND TESTING

Selegey D. (danila-slg@yandex.ru)[1],
Shavrina T. (rybolos@gmail.com)[1],
Selegey V. (Vladimir_S@abbyy.com)[2,3],
Sharoff S. (s.sharoff@leeds.ac.uk)[4]

[1]Moscow State University, Russia
[2]Russian State University for the Humanities, Russia
[3]ABBYY, Russia
[4]University of Leeds, UK

This paper presents a new set of basic tools for morphosyntactic tagging of Russian texts coming from social media. This has been developed within GICR, a project for creating a very large corpus of the Russian-speaking Internet.

This toolset includes a new tagset, obtained via extending and adaptating the tagset proposed by Sharoff et al. It has been tested on a gold standard test corpus of modern social media of about 2 million tokens. A particular feature of our approach is a fully automated process for development of training corpora. Instead of manual annotation we started with the output of the syntactic parser of Compreno. This annotation has been subsequently improved by automatic correction of systematic errors detected through processing of texts from social media. In this paper we show that existing tagging tools (in particular, tnt) produce consistently better results if they are trained with our corpus rather with other available corpora, in particular, those using the disambiguated portion of the Russian National Corpus.

The resulting test corpus is available in open access.

Keywords: automatic morphotagging, morphological tagging, morphotagging for Russian, language of social media

1. Введение

Автоматическая морфоразметка корпусов стала сегодня важной и популярной задачей. Основная причина — желание использовать для лингвистических исследований большие размеченные интернет-корпуса актуального русского языка. Очевидно, что корпуса объемом в десятки миллиардов словоупотреблений практически непригодны для использования в отсутствие разметки, а об их ручной разметке не приходится и говорить. Имеется, однако, несколько серьезных проблем, которые не позволяют пока

получить объективную оценку качества авторазметки и ее пригодности для лингвистики:

1. Разметка корпусов Social Media делается сейчас парсерами, обученными на размеченном подкорпусе с существенно иной жанровой структурой — т. н. «снятнике» НКРЯ [Plungian, 2005] — сравнительно небольшом корпусе с ручной разметкой некоторой смеси текстов НКРЯ не вполне ясной жанрово-тематической принадлежности. Лексические и грамматические особенности этого подкорпуса определили также состав морфословарей и стандарт разметки MSD. Прямой перенос языковой модели на сегменты социальных медиа не позволяет получить морфоразметку нужного качества (Шаров, Беликов et al, 2015).
2. Отсутствует эталонная разметка (золотые стандарты) для оценки качества морфоанализа social media. Первое и последнее тестирование систем морфоразметки проводилось на Диалоге в 2010 году [Liashevskaia O., Astaf'eva I. et al 2010]. Были показаны неплохие результаты, но получены они были на «хороших» текстах. Попытка потестировать «грязные» не показалась тогда особо важной, соответствующая дорожка не собрала участников.
3. Имеются серьезные отличия в подходах к морфоразметке для нужд лингвистических корпусных исследований и для задач автоматического обучения морфопарсеров (POS-тегеров). Эти отличия, однако, явно не сформулированы и никак не учитываются, когда размеченные подкорпуса используются для обучения.

Всё в целом приводит к выводу, что уровень проработки этой проблемы не вполне соответствует ее важности для корпусной лингвистики. Актуальными оказываются две связанные задачи:

1. Создание и обоснование нового тагсета и синхронизированного по грамматической системе морфословаря.
2. Создание достаточно большого тестового корпуса, размеченного в соответствии с этим тагсетом, и выбор технологии его дальнейшего ведения и модификации в условиях большой динамики языка social media.

Необходима также оценка применимости для разметки динамичного и компактного тестового корпуса медленно, но достаточно надежно работающих парсеров «старших» категорий, умеющих снимать грамматическую омонимию на основании полных или локальных синтактико-семантических разборов.

2. Big Social Data и Big Social Corpora

Социальные медиа — такие как блоги (Живой Журнал, LiveInternet, Blogs Mail.ru, etc.), микроблоги (Twitter), социальные сети (VKontakte, Одноклассники, FaceBook, etc.), всевозможные форумы — на сегодняшний момент

являются основным по объему данных[1] источником материала для корпусов актуального русского языка. Язык социальных сетей обладает отличиями в языке и типографике, осложняющими автоматическую разметку текста.

Авторы представляют проект Генерального интернет-корпуса русского языка (ГИКРЯ), сегмент social media которого включает сейчас около 20 млрд словоупотреблений. Результаты первой разметки этого сегмента с помощью стандартно обученных методов оказалась неоднозначными: несмотря на высокую точность определения частей речи и лемматизации (лучшие в сравнении с, например RuTenTen — см. приложение 1), качество снятия омонимии по отдельным категориям оказалось существенно ниже желаемого [Sharoff et al., 2015].

Мысль, что высокие морфо проценты скрывают серьезные проблемы, уже высказывалась в работе [Manning, 2011]. Так, при выборе метрики, оценивающей точность разбора на целых предложениях, у хороших морфопарсеров (например, Standford Part-of-Speech Tagger [Toutanova et al., 2003]) получилась точность в 55–57%.

Использовавшиеся методы оценки (и в особености — точность приписывания полного морфо тега) оказались слабо интерпретируемы с точки зрения определения надежности результатов лингвистических исследований. Необходима дифференциальная оценка с учетом значимости отдельных категорий.

Но прежде чем оценивать точность приписывания морфотегов необходимо разобраться с теми грамматическими категориями, которые они реализуют.

3. Требования к грамматической системе корпусной морфоразметки

При разметке корпусов сегодня используются различные морфопарсеры для русского языка. Наиболее известные — mystem [Segalovich, 2003], Tnt-Russian [Sharoff, Nivre, 2011], Tree-Tagger [Schmid, 1994], Abbyy Compreno [Anisimovich et al 2011], морфопарсер системы ЭТАП.

Все вышеприведенные программы используют разные словари и наборы грамматических категорий: у mystem это тагсет системы ЭТАП [Apresian et al., 2003] и словарь Зализняка [Zalizniak, 1977], у TnT-Russian обычно тагсет системы Multext-east for Russian [Erjavec, 2010] и словарь, полученный автоматически на корпусе со снятой омонимией НКРЯ [Plungian, 2005], у системы Abbyy Compreno — словарь и тагсет собственной разработки, используемые в ряде проектов (Lingvo, FineReader, Compreno). Все эти системы развивались параллельно и результаты их работы серьезно не сравнивались до соревнования на Диалоге в 2010 (в котором наилучшие результаты показала система

[1] Рунет в 2014 г. занимал как минимум 155 экзабайт, или 2,4% данных человечества (http://www.rg.ru/2013/05/14/infa-site.html). К 2020 году прогнозируется рост до 980 экзабайт (2,2% мировых данных). В 2015 году, согласно данным internetworldstats, Россия занимает 6-ое место в мире по количеству интернет-пользователей — 103 147 691 при 70%-ной вовлеченности населения — http://www.cnews.ru/news/top/rossiya_sozdaet_24_mirovogo_obema_dannyh

Compreno). По итогам соревнования был сделан небольшой тестовый корпус на объединенной разметке, но анализ особенностей разметок каждой отдельной системы не проводился.

Чтобы сравнение разных разметок стало возможным, организаторы приняли в 2010 г. несколько упрощающих соглашений, в частности сократили систему частеречных признаков до 6 граммем: существительные, прилагательные, глаголы (в том числе причастия и деепричастия), предлоги, и союзы. Все прочие неизменяемые слова, наречия, вводные слова, частицы, попали в одну категорию. Местоимения и числительные, а также ряд других объектов вовсе не размечались. Соответствующим образом выглядел и получившийся золотой стандарт.

Для корпусной разметки такое решение, объединяющее в одну кучу всё, что Пушкин относил к той самой столь важной для языка «мелкой сволочи»[2], не годится ни с точки зрения обучения парсеров, ни для корпусного исследования.

В современных корпусах наблюдается сейчас некоторое единство в сфере стандартов разметки. Такие проекты, как НКРЯ, RuTenTen, корпус университета Лидс и Araneum Russicum используют стандарт Multext-East for Russian и морфопарсер TnT-Russian. В проекте ГИКРЯ также использовался данный парсер, однако, с модификациями, дающими некоторый выигрыш в качестве на сегменте social media [Sharoff et al., 2015] — см. приложения 1 и 2.

Но проблема «мелкой сволочи» в этой разметке решается непоследовательно, что в значительной степени отражает непоследовательность в разметке самого снятника, в частности при анализе вводных конструкций типа «по большому счету» и т. п.

При формулировании требований к грамматической системе корпусной морфоразметки приходится считаться как с нуждами обучения, так и с потребностями исследователя языка:

1) В интересах обучения парсера грамматическая система должна обеспечивать явное маркирование любых частотных дистрибуционных различий. Только в этом случае возможно эффективное снятие грамматической омонимии.
2) Грамматическая система должна соответствовать запросам пользователей корпуса (ее можно изучать на основании статистики запросов к ГИКРЯ и НКРЯ).
3) Граммемы этих категорий должны быть потенциально «надежно определяемыми» парсерами, иначе их выставление будет лишь сбивать с толку исследователя (такие граммемы могут все же быть полезными, но только при явном понимании пользователем степени их надежности. Это, например, относится к грамматической интерпретации глагольных форм на *-ся*).

---

[2] *…А подбирать союзы да наречья;*
Из мелкой сволочи вербую рать.
Мне рифмы нужны; все готов сберечь я,
Хоть весь словарь; что слог, то и солдат —
Все годны в строй: у нас ведь не парад.
(А. С. Пушкин Домик в Коломне)

Таким образом, граммемы неоднородны с точки зрения их использования:

- часть граммем не нужна для обучения, но требуется исследователю («информационные» категории), например граммемы одушевленности;
- часть граммем может вводиться в систему исключительно для нужд учета специфики дистрибуции высокочастотной лексики замкнутых классов; например, такие псевдограммемы могут быть приписаны формам «нет» в глагольно-предикативном использовании.

На практике приходится также считаться с особенностями парсеров, которые могут быть чувствительны к росту индекса, увеличению числа низкочастотных комбинаций граммем и проч. Желательно, однако, чтобы «архитектурные» ограничения парсера не влияли на разметку золотого стандарта — это вопрос его адаптации к конкретному алгоритму снятия омонимии.

4. Особенности использования грамматической системы синт. парсеров

Кажется разумным использовать разметку парсера, с помощью которого анализировался текстовый корпус. Имеется, однако, проблема, связанная с тем, что парсер Compreno, например, для которого важна полнота морфо- и синтаксических описаний, использует большое число конвенций, являющихся по сути артефактами конкретной модели описания. Кроме того, необходимость описания синтаксических явлений порождает большой класс синтаксических граммем, которые «отменяют» морфологические, возникают проформы разных типов и т. д.

Это делает мэппинг грамматической системы Compreno и любой другой системы полного анализа на новый стандарт морфоразметки достаточно сложным делом, но зато позволяет в дальнейшем использовать эту «тяжелую» разметку для новых тестовых подкорпусов.

Нельзя не коснуться вопроса о перспективах полной синтаксической разметки больших корпусов социальных медиа. Анализ разборов Compreno показывает что качество построения полных семантико-синтаксических структур на специфических текстах этого сегмента все же заметно хуже, чем результаты на текстах non-fiction, на которых, например, в условиях тестирования 2012 года были показаны очень высокие результаты [Toldova S., Sokolova E. et al. 2012]. Кроме того, ресурсы на такую разметку оказались бы слишком велики (не говоря уже о вопросах некоммерческого использования этих технологий).

При этом статистика запросов к ГИКРЯ показывает, что для лингвистических исследований на мегакорпусах, сильно смещенных в сторону лексических явлений, гораздо чаще не нужен полный синтаксический разбор, и при хорошем качестве снятия омонимии можно использовать локальные Sketch-грамматики.

Таким образом, в данной работе мы используем морфословарь и синтаксическую разметку системы Compreno (от которой берем по соглашению CBR только снятую грамматическую омонимию). Оправданность такого решения хорошо видна при сравнении качества снятия омонимии на одних и тех же

текстах социальных медиа в разметке ГИКРЯ сейчас (TNT-парсер) и в разметке Compreno (после мэппинга, но до оптимизации). Данные см. в таблице 1.

Таблица 1. Точность автоматического снятия омонимии на некоторых граммемах имени существительного

граммема	Точность при разметке TnT	Точность при разметке Compreno
Неодушевленный номинатив	0,792	0,947
Неодушевленный аккузатив	0,858	0,884
Одушевленный аккузатив	0,661	0,980
Одушевленный генитив	0,890	0,890
Субстантивы (на выборке из омонимичных вхождений)	0,680	0,916
Прилагательные (на выборке из омонимичных вхождений)	0,900	0,918

5. Принципиальные проблемы автоматической морфоразметки social media

При автоматическом создании текстового корпуса приходится решать 2 проблемы:

1. «Мэппинг» — уже описанная выше задача мэппирования грамматической системы используемого «тяжелого» парсера на новый морфостандарт;
2. «Оптимизация» — исправление частотных систематических ошибок, допускаемых этим парсером.

Несмотря на высокую точность, заметно превосходящую результаты, получаемые обычными тегерами, в непосредственно полученном после мэппирования корпусе есть еще, с чем можно побороться. Речь идет о систематических ошибках, которые исправляются полностью автоматически в любом новом цикле «разбор парсером — мэппирование — финальное улучшение». Примерные результаты разбора после мэппирования (они незначительно зависят от разбираемого подкорпуса социальных медиа) в Таблице 2:

Таблица 2

Формат	Среднее кол-во граммем на слово	Точность разметки
Abbyy Compreno	7	ABBYY 0,942
MSD	4	TnT-Russian 0,887
ЭТАП	4	Mystem 0,927

Статистика остаточных ошибок по типам в GICRMorpho и ГИКРЯ в таблице 3:

Таблица 3

неправильная лемма	38,18%
неправильная часть речи	23,64%
падежная омонимия	23,64%
неправильная собственность	5,45%
неправильная транзитивность	5,45%
неправильная одушевленность	1,82%
омонимия формы	1,82%

Большая часть ошибок (ошибки на лемму и часть речи) вызваны опечатками, сленгом и именами собственными, неизвестными словарю (более подробно — см. Таблицу 4).

Таблица 4

Причина ошибки	**Штук**	**Доля**
опечатка	7	17,50%
сленг	14	35%
неизменяемое слово	1	2,50%
имя собственное	3	7,50%
омонимия	15	37,50%

6. Данные о новом тестовом корпусе GICRMorth

Тестовый корпус Social Media представляет собой выборку текстов из Живого Журнала объемом в 2 млн слов. Данные тексты были очищены от html-разметки и затем направлены на анализ в систему ABBYY Compreno. Полученный материал был очищен от синтаксической разметки и смэппирован в новый тагсет MSD-GICR. Из обучающей выборки были извлечены триграммные частоты.

На данном этапе распространяемый по лицензии корпус следует рассматривать как тестовый: безусловно, по результатам тестирования можно ожидать как изменений в тагесете, так и в разметке и в самом подборе текстов. Важно, что методика получения корпуса является полностью автоматической и может быть повторена с другой выборкой данных (например, для другого сегмента Social Media).

7. Результаты, открытые вопросы

Основные результаты нашей работы сводятся к двум важным составляющим: первая — это первый золотой стандарт морфологической разметки в 2 миллиона словоформ, обладающий высоким качеством морфологической разметки и являющийся общедоступным источником, на котором другие исследователи смогут обучать свои морфологические и синтаксические парсеры, ориентированные на обработку текстов social media.

Пример разметки в новом тагсете приведен в таблице 5:

Таблица 5

Vertical text	**Lemma**	**MSD-GICR**	
Если	[если]	C	#союз
хочешь	[хотеть]	V-ip2s-a-p-ym	#глагол, личная форма, наст.вр., 2-е лицо, ед. ч., активный залог, несов., перех.
тусить	[тусить]	V-n----a-p-nm	#глагол, инфин., активный залог, несов., неперех.
—			
туси	[тусить]	V-m-2s-a-p-nm	#глагол, пов.накл., 2-е лицо, ед. ч., активный залог, несов., неперех.
.			
Если	[если]	C	
хочешь	[хотеть]	V-ip2s-a-p-ym	
бухнуть	[бухнуть]	V-n----a-e-ym	#глагол, пов.накл., 2-е лицо, ед. ч., активный залог, соверш., перех.
—			
бухни	[бухнуть]	V-m-2s-a-e-ym	#глагол, пов.накл., 2-е лицо, ед. ч., активный залог, соверш., перех.
.			

Вторая составляющая — это n-граммная статистика для POS-тегера, которая будет применена к соответствующим сегментам ГИКРЯ — социальным сетям и блогам.

Качество разметки нового парсера в сравнении с качеством стандарта — в таблице 6:

Таблица 6. Итоговое качество частеречной разметки

Часть речи	**Точность**	**Полнота**	**F-мера**
1. Существительное	0,989	0,960	0,974
2. Глагол	0,995	0,979	0,987
3. Прилагательное	0,978	0,978	0,978
4. Местоимение	1,000	0,896	0,945
5. Наречие	0,940	0,935	0,937
6. Предлог	1,000	0,972	0,986
7. Союз	0,927	0,962	0,944
8. Числительное	0,957	0,978	0,967
9. Частица	0,964	0,891	0,926
10. Междометие	1,000	0,585	0,738
11. Предикатив	1,000	0,900	0,947
12. Вводное слово	0,954	0,807	0,874
Макроусреднение	**0,975**	**0,904**	**0,934**

Анализ ошибок результирующей разметки показал, что основная масса ошибок приходится на:

- Омонимию разрядов местоимений («*его*»-«*его*»)
- Омонимию субстантивов и причастий («*данные*»)
- Омонимию форм у неизменяемых существительных («бананы, груши, *манго*»)
- Омонимию форм частиц и союзов («*однако*»)
- Опечатки
- Неправильную лемматизацию несловарных слов

Открытыми остаются вопросы о включении в цепочку морфологической обработки блока псевдолемматизации и исправления орфографии. В дальнейшей работе над морфологической разметкой мы планируем включить наработки из работы [Sorokin, Khomchenkova, 2016], которые позволят улучшить качество обработки несловарных слов, а также результаты работы [Sorokin, Shavrina, 2016], позволяющие стабилизировать качество разметки независимо от грамотности авторов текстов и тем самым избежать сдвигов статистики употреблений отдельных слов.

В процессе разработки мы создали правила меппинга, позволяющие перекодировать существующие форматы MSD, MSD-GICR [Sharoff et al., 2015], mystem и ABBYY Compreno в новый формат разметки, доступный исследователям.

8. Заключение

В данной статье мы представляем результаты создания нового обучающего корпуса и нового стандарта морфоразметки social media для русского языка.. Модель обучения на триграммах, представленная в [Sharoff, Nivre, 2011] была

воспроизведена на основе новой обучающей выборки. В рамках этой работы результаты автоматической морфологической разметки ABBYY Compreno были перенесены на новый тагсет, который представляет из себя дальнейшее развитие Multext. Полученный морфотаггер показывает возможность успешного снятия морфологической омонимии и лемматизации на материале социальных сетей.

Этот морфотаггер будет использован для разметки и развития в рамках проекта «Генеральный интернет-корпус русского языка». Подкорпус из 2 млн словоупотреблений social media, размеченный в новом наборе категорий с автоматически снятой омонимией, доступен для свободного использования [http://www.webcorpora.ru/news/282]. Авторы надеются, что их разработка поможет NLP-community в развитии и обучении их морфологических и синтаксических парсеров на базе нового стандарта разметки.

Благодарности

Авторы выражают благодарность компании ABBYY за возможность использования парсера Compreno в рамках академической лицензии CBR. Мы благодарим Анастасию Сиротину, чьи комментарии были крайне полезны при анализе морфоразметки. Выражаем особую благодарность Валерию Новицкому и Андрею Андрианову из ABBYY, чья помощь и консультации сделали возможными морфологические эксперименты, описанные в статье.

Литература

1. *Anisimovich K. V., Druzhkin K. Ju., Minlos F. R., Petrova M. A., Selegey V. P. Zuev K. A.* Syntactic and semantic parser based on ABBYY Compreno linguistic technologies. In: Computational linguistics and intellectual technologies. 2012, Vol. 11. http://www.dialog-21.ru/digests/dialog2012/materials/pdf/Anisimovich.pdf
2. *Apresian J., Boguslavskii I., Iomdin L., Lazurskii A., Sannikov V., Sizov V., Tsinman L.* (2003) ETAP-3 Linguistic Processor: a Full-fledged NLP Implementation of the MTT. First International Conference on Meaning-Text Theory: 279–288.
3. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.* (2013), Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. In Proc. Web as Corpus Workshop (WAC-8).
4. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.* (2013), Corpus as language: from scalability to register variation. In Proc. Dialogue, Russian International Conference on Computational Linguistics, Bekasovo.
5. *Dimitrova, L., T. Erjavec, N. Ide, H.-J. Kaalep, V. Petkevic, and D. Tufis* (1998), MULTEXT-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In COLING-ACL '98. Montreal, Quebec, Canada.
6. *Erjavec T.* (2010) Multext-east Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10).

7. *Gareyshina A., Ionov M., Lyashevskaya O., Privoznov D., Sokolova E., Toldova S.* (2012) RU-EVAL-2012: Evaluating Dependency Parsers for Russian. Proceedings of COLING 2012: Posters. P. 349–360. URL: http://www.aclweb.org/anthology/C12–2035.
8. *Jongejan B. and Dalianis H.* (2009) Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Suntec, Singapore: Association for Computational Linguistics, 2009. s. 145–153
9. *Kilgarriff A., Baisa V., Busta J., Jakubicek M., Kovar V., Michelfeit J., Rychly P., Suchomel V.* (2014). "The Sketch Engine: ten years on". Lexicography (Springer Berlin Heidelberg) 1 (1): 7–36.
10. *Kilgarriff A., Rychly P., Smrz P., Tugwell D.* (2004) The Sketch Engine/ Proc. Euralex. Lorient, France
11. *Liashevskaia O., Astaf'eva I., Bonch-Osmolovskaia A., Gareishina A., Iu., G., D'iachkov V., Ionov M., Koroleva A., Kudrinski M., Litiagina A., Luchina E., Sidorova E., Toldova S., Savchuk S., and Koval' S.* (2010) Evaluation of Automatic Text Parsing Methods: Morphological Parsers in Russian [Otsenka Metodov Avtomaticheskogo Analiza Teksta: Morfologicheskie Parsery Russkogo Iazyka]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2010" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2010"): 318–326.
12. *Manning C. D.* (2011), Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In CICLing Conference on Intelligent Text Processing and Computational Linguistics.
13. *Plungian V. A.* (2005) What do We Need Russian National Corpus for? [Zachem Nuzhen Natsionalnii Korpus Russkogo Iazyka?] Natsionalnii Korpus Russkogo Iazyka: 6–20.
14. *Schafer R.* (2015) FYI: COW: Free, Large Web Corpora in European Languages. LINGUIST List 26.2114 (web resource: http://linguistlist.org/issues/26/26-2114.html)
15. *Schmid H.* (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.
16. *Segalovich I.* (2003) A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine.
17. *Sharoff S. A., Belikov V. I., Kopylov N. Y., Sorokin A. A., Shavrina T. O.* (2015) Corpus with Automatically Resolved Morphological Ambiguity: to the Methodology of Linguistic Research. In Dialogue, Russian International Conference on Computational Linguistics, Moscow.
18. *Sharoff S., Nivre J.* (2011) The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2011"], Bekasovo, pp. 591–605.

19. *Shavrina T., Sorokin A.* (2015) Modeling Advanced Lemmatization for Russian Language Using TnT-Russian Morphological Parser. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2015"
20. *Sokirko A.* (2004) Morphological Modules on the web-site www.aot. ru [Morphologicheskie Moduli na saite www.aot.ru]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2004" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2004").
21. *Sokirko A., Toldova S.* (2005) Sravnenie Effektivnosti Dvukh Metodik Sniatiia Lexicheskoi i Morfologicheskoi Neodno znachnosti dlia Russkogo Iazyka. Internet-matematika.
22. *Sorokin A. A., Baitin A. V., Galinskaya I. E., Shavrina T. O.* (2016) SpellRuEval: The First Competition On Automatic Spelling Correction For Russian. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2016"
23. *Sorokin A. A., Shavrina T. O.* (2016) Automatic spelling correction for Russian social media texts. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2016"
24. *Sorokin A. A., Khomchenkova I. A.* (2016) Automatic detection of morphological paradigms using corpora information.
25. *Toldova S., Sokolova E. et al.* NLP evaluation 2011–2012: Russian syntactic parsers. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012"
26. *Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer.* (2003) Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252–259.
27. *Zalizniak A.* (1977) Russian Grammar Dictionary [Grammaticheskii Slovar' Russkogo Iazyka. Russki Iazyk].

Приложения

Приложение 1 — Table 3 from [Sharoff et al., 2015, 8]

RuTenTen			**GICR**		
Part of Speech	**Precision**	**Recall**	**Part of Speech**	**Precision**	**Recall**
1. Noun	0,948	0,987	1. Noun	0,997	0,99
2. Verb	0,966	0,976	2. Verb	0,998	0,998
3. Adjective	0,942	0,969	3. Adjective	0,953	0,997
4. Pronoun	0,988	0,975	4. Pronoun	1,000	1,000
5. Adverb	0,927	0,914	5. Adverb	0,974	0,913
6. Preposition	1,000	0,997	6. Preposition	1,000	0,998
7. Conjunction	0,993	0,991	7. Conjunction	0,993	0,993
8. Numeral	0,797	0,911	8. Numeral	0,98	1,000
9. Particle	0,986	0,983	9. Particle	0,996	0,996
10. Inetrjection	1,000	0,551	10. Inetrjection	1,000	0,9
11. Other	0	0	11. Predicative	1,000	0,81
12. Abbreviation	0	0			
Microaverage:	0,979	0,975	Microaverage:	0,99	0,963
Macroaverage:	0,7956	0,7712			
Macroaverage without 11–12:	0,955	0,925	Macroaverage:	0,991	0,99

Приложение 2 — Table 2 from [Sharoff et al., 2015, 7]

Old morphology of GICR			**New morphology of GICR**		
Part os speech	**Precision**	**Recall**	**Part os speech**	**Precision**	**Recall**
1. Noun	0,992	0,987	1. Noun	0,997	0,99
2. Verb	0,989	0,991	2. Verb	0,998	0,998
3. Adjective	0,95	0,943	3. Adjective	0,953	0,997
4. Pronoun	0,997	0,997	4. Pronoun	1	1
5. Adverb	0,808	0,947	5. Adverb	0,974	0,913
6. Preposition	1	1	6. Preposition	1	0,998
7. Conjunction	0,979	0,996	7. Conjunction	0,993	0,993
8. Numeral	0,815	0,963	8. Numeral	0,98	1
9. Particle	0,996	0,959	9. Particle	0,996	0,996
10. Inetrjection	1	0,714	10. Inetrjection	1	0,9
11. Other	0	0	11. Predicative	1	0,81
12. Abbreviation	0	0			
Microaverage:	**0,794**	**0,796**	**Microaverage:**	**0,99**	**0,963**
Macroaverage:	**0,976**	**0,987**	**Macroaverage:**	**0,991**	**0,99**
Macroaverage without 11–12:	**0,953**	**0,95**			

Приложение 3

Руководствуясь приведенными выше принципами, мы создали итоговый тагсет, основанный на позиционной системе MSD, но качественно дополненный граммемами Abbyy Compreno, помогающими снимать омонимию, и лишенный граммем, сложных в определении и отягчающих реализацию разметки:

Position	Code	Meaning
0	N	Noun
1	p/c	Type: proper/common
2	m/f/n/c/-	Gender: Masculine/Feminine/Neuter/Common/Undefined (for pluralia tantum)
3	s/p	Number: Singular/Plural
4	n/g/d/a/l/i/v	Case:Nominative/(Genitive|Partitive)/Dative/Accusative/(Locative|Prepositional)/Instrumental/Vocative
5	n/y	Animatedness: Inanimate/Animate
6	p/l/-	Case2: Partitive/Locative/(Nominative|Genitive|Dative|Accusative|Prepositional|Instrumental|Vocative)
0	V	Verb
1	-	
2	i/m/n/p/g/x	GrammaticalType: Indicative/Imperative/Infinitive/Participle/Adverb/WordNet
3	p/f/s/-/*	Tense: present/future/past/Undefined/Cannot be disambgued
4	1/2/3/-	Person: First/Second/Third/Undefined
5	s/p/-	Number: Singular/Plural/Undefined
6	m/f/n/-	Gender: Masculine/Feminine/Neuter/Undefined
7	a/p/s	Voice: Active/Passive/VoiceSya
8	s/f/-	ParticipleShortness: ShortForm/FullForm/Undefined
9	p/e/-/*	Aspect: Imperfective/Perfective/Undefined/Cannot be disambgued
9	n/g/d/a/l/i/-	Case: Nominative/Genitive/Dative/Accusative/Prepositional/Instrumental/Undefined
10	n/y	Transitivity: Intransitive/Transitive
11	m/b	Pairness: (MonoAspectual|Paired)/BiAspectual
0	A	Adjective
1	-	
2	p/c/s	DegreeOfComparison: Positive/Comparative/Superlative
3	m/f/n/-	Gender::Masculine/Feminine/Neuter/Undefined
4	s/p	Number: Singular/Plural/
5	n/g/d/a/l/i	Case: Nominative/Genitive/Dative/Accusative/Prepositional/Instrumental
6	s/f	AdjectiveShortness: ShortForm/FullForm

Position	Code	Meaning
0	P	Pronoun
1	p/d/i/s/q/x/z/n	ReferenceClass::RCPersonal/RCDemonstrative/RCIndefinite/RCPossessive/RCInterrogative/RCReflexive/RCNegative/RCAttributive
2	1/2/3/-	Person: First/Second/Third/Undefined
3	m/f/n/-/*	Gender: Masculine/Feminine/Neuter/Undefined/»все» and «всё» cannot be disambgued
4	s/p/-/*	Number: Singular/Plural/Undefined/»все» and «всё» cannot be disambgued
5	n/g/d/a/l/i	Case: Nominative/Genitive/Dative/Accusative/Prepositional/Instrumental
6	n/a/r	Syntactic Type: Nominal/adjectival/adverbial
7	s/f/-	Shortnesss: ShortForm/FullForm/Undefined
0	R	Adverb
1	p/c/s	DegreeOfComparison: Positive/Comparative/Superlative
0	W	Predicative
0	S	Preposition
1	p	Type: preposition
2	-	
3	g/d/a/l/i	Case: Genitive/Dative/Accusative/Prepositional/Instrumental
0	C	Conjunction
0	M	Numeral
1	c/l/o	Type: cardinal/collect/ordinal
2	m/f/n/-	Gender: Masculine/Feminine/Neuter/Undefined
3	s/p/-	Number: Singular/Plural/Undefined
4	n/g/d/a/l/i	Case: Nominative/Genitive/Dative/Accusative/Prepositional/Instrumental
5	l/d/r	Form: numeral/arabic digit/roman digit
0	Q	Particle
0	I	Interjection
0	H	Parenthetical phrase
0	X	Residual

ДИСКУРСИВНЫЕ СЛОВА И КОММУНИКАТИВЫ

Шаронов И. А. (Igor_sharonov.@mail.ru)

Российский государственный гуманитарный университет, Москва, Россия

Ключевые слова: коммуникативы, диалог, ответные реплики, дискурсивные слова, прагматика, лексикография

DISCURSIVE WORDS AND COMMUNICATIVES

Sharonov I. A. (Igor_sharonov.@mail.ru)

RSUH, Moscow, Russia

The article is devoted to communicatives — special use of words, idioms and short sentences in dialogical positions of stereotype responses (response particles), intended to agree, disagree, answer some etiquette formulae, or to express different emotions. These are conversational formulae like *Da; Net uzh; Kakoe tam; Obaldet'!; Na zdorovje!*, etc. Communicatives consist of particles and idiomatic constructions; they are semantically empty and pragmatically specific. These units are regularly used in conversation, but not a single Russian dictionary has yet seen light where one could get compete information about communicatives and their occurrence in conversation. Only very few communicatives can be found in explanatory dictionaries and dictionaries of idioms. But their description in those rare cases is limited by their intention (affirmative response, doubt, etc.) or their function — either an etiquette answer (used to express thanks, regret) or an emotional response (used to express surprise, joy, grief). In some cases, communicatives may be marked as synonyms. Some words and idioms may function both as discourse words and communicatives, and some modern dictionaries claim to contain full information concerning their semantics and use. But the attention in the dictionaries is focused mostly on narrative, not dialogical, contexts, which distorts communicatives' actual use. The objective of the article is to compare characteristics of discursive words and communicatives. In the few examples we try to demonstrate the differences between the meaning and usage of these units and argue for the compiling the special Dictionary of Communicatives.

Key words: communicatives, dialog, responses (response particles), discursive words, pragmatics, lexicography

1. Коммуникативы и их место в словарях

Коммуникативы — это особые употребления слов, фразем и коротких предложений в позиции ответных реплик диалога для стереотипного выражения оценки, мнения и эмоции как реакции на высказывание собеседника (напр.: *Да; Нет уж; Какое там; Обалдеть!; На здоровье!* и др.). В силу частотности и стереотипности употребления в речи такие реплики крайне лаконичны. Они состоят в основном из частиц, вводных слов и знаменательной лексики, теряющей в «коммуникативных» употреблениях лексическое значение. Традиционные лингвистические теории обычно выводят стереотипные ответные реплики за рамки описания, включают их в размытый класс междометий[1].

Некоторое количество слов, используемых в речи в качестве коммуникативов, можно обнаружить в толковых и специальных словарях (см. напр., Ожегов 1990, Рогожникова 1983, Квеселевич, Сасина 1990, СССРЯ 1997, Шимчук, Щур 1999 и др.). В толковых словарях коммуникативы встречаются в словарных статьях **ДА**, **НЕТ**, **ВОТ**, **ЧТО**, **КАК**, **НУ**. Неоднословные сочетания в качестве коммуникативов иногда можно найти в словарных статьях их ключевого компонента «за ромбом», в зоне фразеологии. Так, в Ожегов 1990 коммуникатив *А то* находится в словарной статье **А**, *А то нет* — в статье частицы **НЕТ**, *Как же!* — в словарной статье **КАК**, *На тебе* (*нате вам*) в статье **НА**. Идиоматические сочетания, способные выполнять функцию коммуникатива, можно также найти во фразеологических словарях (Молотков 1986, Lubensky 1995, Баранов, Добровольский 2009 и др.). В целом же приходится признать, что общее число собранных в словарях коммуникативов составляет лишь небольшой процент реально функционирующих в диалогической речи стереотипных реплик, выполняющих функцию ответного речевого акта.

Причина слабого внимания лексикографов к коммуникативам заключается в том, что традиционно иллюстративный материал для словарей — нарративные тексты. Частотность диалогических употреблений языковой единицы в таких текстах значительно уступает нарративным. Сравните общую статистику употреблений нескольких коммуникативов в сравнении с прочими употреблениями соответствующих единиц по материалам НКРЯ на ноябрь 2014 г.[2]:

Яз. единица	Коммуникатив	Другие употребления	Соотношение
• Да уж	• 423 вхождения	• 3 335 вхождений.	**1/7**
• Куда там	• 72 вхождения	• 534 вхождения	**1/7,41**
• Ну уж	• 189 вхождений	• 2 008 вхождений	**1/10,6**
• *Во*	• 445 вхождений	• 311 103 вхождения.	**1/700**
• Неужели	• 20 вхождений	• 22 226 вхождений	**1/1111,3**

[1] Обзор работ, посвященных языковому статусу коммуникативов (или релятивов) см. в Шаронов 1996, Колокольцева 2001).

[2] Методика выявления «коммуникативных» употреблений описана в Шаронов 2015.

Время поиска необходимого материала значительно сокращается, благодаря разметке в НКРЯ[3].

2. Традиционные толкования коммуникативов

Толкуются представленные в словарях коммуникативы обычно либо через синонимы, либо через указание на тип речевого акта и на эмоцию: *согласие/несогласие, разрешение, возражение, запрещение, ответ на благодарность, радость, удивление* и т. д. Такого рода толкования могут быть полезны для понимания репликового слова только при аудировании и чтении. Современные семантические теории и их лексикографические реализации предлагают интегральное, «активное» толкование слова, необходимое не только для рецептивных, но и для продуктивных видов речевой деятельности — говорения и письма (см. НОССРЯ 1997, 2000, 2003, АСРЯ 2014 и др.). Интегральные толкования призваны помочь пользователю словаря не только понимать, но и употреблять языковую единицу в своей речи. На «активность» претендуют также теории анализа дискурсивных слов (дискурсивов), служащих для выражения коммуникативно-прагматической информации в высказывании (ПДС 1993, ДСРЯ 1998).

По своему лексическому составу коммуникативы в значительной степени совпадают с дискурсивами: частицами, модальными наречиями и вводными словами. В процессе исследования дискурсивов, предпринятых в рамках Московской семантической школы, был выработан арсенал аналитических приемов, позволяющих выявлять указательные, выделительные, усилительные, анафорические и другие функции этих единиц (см. Яковлева 1994; Баранов, Кобозева 1988; Разлогова 1996 *и многие др.*). В этих и близким им исследованиях дискурсивные и коммуникативные употребления языковой единицы описываются вместе как нечто единое, в словарях — в рамках одной словарной статьи и под одним толкованием. При этом диспропорция в описании бросается в глаза: богатая палитра выделенных дискурсивных характеристик единицы резко контрастирует с лаконизмом ее «коммуникативных» свойств. Причина — в акценте исследования на поведении слова в предложении, выявлении его роли как модификатора смысла предложения. Для адекватного описания коммуникатива арсенал приемов, выработанный в рамках дискурсивного анализа, оказывается мало приспособленным.

---

[3] Все же наиболее адекватными для описания коммуникативов были бы представительные корпуса устной русской диалогической речи, но их, к сожалению, пока нет (по крайней мере, в открытом доступе).

3. Лингвистические исследования функционирования коммуникативов

В 90-х годах XX века на волне интереса к семантике русских частиц было защищено несколько диссертаций о способах описания частиц со значением согласия и возражения (верификативных высказываний) [Добрушина 1993, Галактионова 1995, Михайлова 1999]. В этих работах были выявлены важные методологические положения, определяющие коммуникативный статус, границы и условия употребления ряда ответных реплик-частиц в зависимости от типа предшествующего высказывания. В более поздних исследованиях, непосредственно посвященных ответным репликам, также отмечается неразрывная связь с прототипической стимулирующей репликой, ее интенциональных, синтаксических, семантических и прагматических характеристик [см. Sorjonen 2002].

Конкретные описания форм зависимости семантики и прагматики ответных реплик от иллокутивных и формальных особенностей реплик-стимулов представлены, в частности, в работах по анализу бытового диалога[4].

4. Разделение дискурсивов и коммуникативов

Описание единицы в дискурсивной функции, выполняя которую слово выступает в качестве модификатора высказывания, и в репликовой функции, где языковая единица выполняет самостоятельный ответный речевой акт, необходимо проводить раздельно и в разных форматах. Можно говорить о двух основных причинах для этого.

4.1. Семантика единицы в дискурсивной и в репликовой функции может различаться: дискурсивные компоненты значения свойства ее «выветриваются», интенциональные значения меняются.

Возьмем в качестве примера фраземы *сделай милость (одолжение)* и *будь так добр.* Эти единицы традиционно описываются как показатели (деликатной) просьбы, хотя такое описание не совсем точно — фраземы в качестве дискурсивов только **вводит** в текст деликатную просьбу, а **выражает** ее последующая предикация:

— *Сделай милость,* принеси/сходи/открой…

— *Будь так добр,* расскажи/позвони/узнай…

В качестве коммуникативов данные фраземы реализуют самостоятельный речевой акт: благодарное, возможно, несколько высокопарное **согласие** в ответ **на любезное предложение** собеседника, на выражение его готовности сделать что-л. необходимое, полезное, приятное для говорящего или для третьих лиц:

(1) — *Зернистой икорочки? —* ***Будьте добры.*** [А. Мариенгоф. Циники]

---

[4] Подробнее об этих работах см. в Sorjonen 2001, Paukkeri 2006.

(2) — *Пойду оформлю документы на ваше освобождение.*
— ***Будьте так любезны***. [А. Ростовский. Русский синдикат]

(3) — *Я им займусь, — торопливо заверил мэра подошедший вслед Смирнов.*
— ***Будь так добр***. [С. Таранов. Черт за спиной]

В работе Кустова [2012] рассматривается обособление некоторых глагольных форм от парадигмы глагола в условиях ведения диалога, переход их в коммуникативы (в терминологии автора — иллокутивы), а потом, следующим «эволюционным» этапом — в дискурсивы. Нам последовательность перехода представляется обратной: коммуникативы следуют за дискурсивами. Если рассматривать эволюцию с точки зрения потерь функциональных возможностей знаменательного слова (парадигмы, синтаксических связей и т. п.), коммуникативы находятся дальше, они лингвистически «беднее» чем дискурсивы. Косвенным аргументом нашему утверждению может служить также история формы *положим*. Данная форма традиционно используется в русском языке как дискурсив, вводное слово, выражающее уступительное допущение[5]. Ср. примеры из XIX, XX и XXI вв:

(4) *«Хорошо, — продолжал он, —* ***положим****, что тогда действительно… Было выпито — это так.* [М. Е. Салтыков-Щедрин. Пошехонские рассказы (1883–1884)]

(5) ***Положим****, Георгий Николаевич скажет Корнилову «молчи», а Корнилов его не послушает, тогда что?* [Ю. О. Домбровский. Факультет ненужных вещей (1978)]

(6) — *Ну, этого вы,* ***положим****, гарантировать не можете, — недобро усмехнулся Разумов.* [Александра Маринина. Последний рассвет (2013)]

В первой половине XX века появились (а к середине века исчезли) репликовые употребления *положим* со значением насмешливого возражения. Ср.:

(7) — *В таких делах не позорно не сдержать слова… Любовь… это… сам знаешь, совсем особенная штука… —* ***Ну, положим****… Особенная. Так извини, Коля, а в тебе только мужское самолюбие говорит. Втюрился, так и в тебя должны втюриться.* [К. М. Станюкович. Пари (1901)]

---

[5] МАС, Т. 3 в статье ПОЛОЖИТЬ дает следующие значения: **4.** *1 л мн ч буд вр* **положим** *в зн. вводного сл.* Предположим, будем считать возможным, допустим. **5.** *1 л мн ч буд вр* **положим** *в зн. вводного сл. (при противопоставлении, обычно с последующим* «но»). Употребляется при выражении уступительного допущения в значении «пускай», «хотя».

(8) *— У меня больше денег, чем у твоего Котьки. <…> —* ***Ну, положим****, — протянула Галя. — И совсем не больше. Котька знаешь сколько у медника получает?* [В. Беляев. Старая крепость (1937–1940)]

(9) *— Да, — ответил я. — Я даже вижу каналы. Я знал, что на Марсе живут люди — марсиане — и что они выкопали неизвестно для чего на своей планете громадные каналы. —* ***Ну, положим****! — сказал отец. — Не выдумывай! Никаких каналов ты не видишь.* [К. Паустовский. Золотая роза (1955)]

4.2. Теряя дискурсивные характеристики, коммуникативы одновременно «обрастают» новыми свойствами: согласованием по иллокутивной функции [см. Падучева 1982] и экспрессивными характеристиками (эмоциональными и оценочными, выражающимися в устной коммуникации мимикой, жестами и относительно фиксированной интонацией), а также возрастными, социальными и стилистическими ограничениями на употребление единицы.

Завершим доклад рассмотрением частицы *именно* в дискурсивно ориентированных словарях и в материалах Словаря коммуникативов, находящемся в настоящее время в процессе составления.

Описание частицы *именно* с опорой на дискурсивные значения даются в ПДС 1993 и в НОССРЯ 2000. В обоих словарях отмечается, что использование частицы предполагает наличие ряда возможных ситуаций, объектов или наименований; выбор говорящего конкретной ситуации (объекта, наименования) маркируется выделительной частицей для противопоставления всем прочим возможным ситуациям (объектам, наименованиям). Словарные статьи довольно подробно анализируют контексты частицы *именно* в разных синтаксических позициях предложения, а также описывают использование *именно* в качестве ответной реплики. Все употребления — дискурсивные и репликовые — авторы стремятся представить как вариации общего значения, которое в целом определяется через акцент на важности или точности некоего Р, при котором стоит частица[6].

Смена «фокуса» описания на репликовые употребления позволяют выявить два коммуникатива — *Именно так* и *Вот именно*, используемые в ответ на разные типы инициирующих реплик собеседника и имеющие каждый свое значение. Коммуникатив *именно* и его более экспрессивный дублированный вариант оказываются лишь неполными реализациями этих фразем. В НКРЯ можно обнаружить более 20 репликовых употреблений для каждой фраземы и их неполных реализаций, что позволяет выявить их типовые контексты употребления.

---

[6] В СДС репликовое употребление *Вот именно* определяется парафразой: «В самую точку» и оценивается как «подходящее» с точки зрения говорящего. Вводится представление о наличии альтернативных возможностей, на фоне которых происходит выбор. В НОССРЯ *Вот именно* в репликовой функции толкуется как «энергичное согласие с тем, что сказал собеседник».

Именно так используется обычно после **предположения, догадки** собеседника, выраженной в форме общего вопроса. Коммуникатив служит для убедительного подтверждения правильности гипотезы (исходного предположения вопроса). Используется такая ответная реплика чаще в формальных ситуациях, говорящий выступает как эксперт. Подтверждая, он может сделать глубокий кивок головой.

(10) — *Позвольте спросить вас, <…> разговор, кажется, идёт об Элен Грен, артистке театра? —* ***Именно так****, — ответил пассажир, оскаливаясь с фальшивой любезностью человека, чувствующего своё превосходство.* [А. С. Грин. Ива]

(11) — *То есть вы хотели бы вернуть детям, а, возможно и их родителям, тот самый слух к чужой боли? —* ***Именно так****.* [А. Гулина. Слух к чужой боли]

(12) *Процентов пять, а? Или, может — восемь?.. <…> — Да, товарищ Сталин, — убеждённо подтвердил Абакумов. —****Именно так****, процентов пять. Или семь.* [А. Солженицын. В круге первом]

Возможны, хотя встречаются значительно реже, гипотетические утверждения, так же требующие от говорящего категорического подтверждения.

(13) — *К чему агитировать Виктора Павловича? У него сердце русского советского патриота, как и у всех нас. — Конечно, — сказал Шишаков, —* ***именно так****.* [В. Гроссман. Жизнь и судьба]

(14) — *Белая мантия, красный подбой! Понимаю! —восклицал Иван. —* ***Именно так!*** [М. А. Булгаков. Мастер и Маргарита]

Отличие таких гипотетических утверждений от прочих — в возможности использования сразу после них вопросительных частиц *(ведь) правда? точно? ну?* и т. д.

Коммуникативы *Именно* и *Именно, именно* используются в аналогичных контекстах, являясь неполными реализациями фраземы. Ср.:

(15) *Ради этих голосов политтехнологи и устроили «ЮКОС»— шоу? —* ***Именно*** (= Именно так). [Е. Костюк. Сверхбедные против свербогатых (2003) // «Время МН», 2003.07.30]

(16) — *Вам трехразовое питание, море, хорошую погоду и чтобы убирали в доме? — спросила та тетя. —* ***Именно, именно*** (= Именно так!). [Л. Петрушевская. Город Света]

Коммуникатив *Вот именно1* используется в ответ на мнение собеседника — его **оценочное суждение**, **утверждение, вывод,** выраженное в форме

утвердительного высказывания. Коммуникатив выражает не просто согласие, а горячее одобрение, солидарность с собеседником в оценке объекта. Говорящий может слегка выдвинуть подбородок вперед, направить указательный палец на собеседника.

(17) — *А по-моему, какой канал ни включишь, везде криминал.* — ***Вот именно***. [Ю. Кантор. Сергей Безруков: «На телевидении я делаю то, что хочу» (2001) // «Известия», 2001.09.24]

(18) — *Ничего себе шуточки…*— *пробормотал я.* — ***Вот именно!*** *Они не похожи на шутников.* [В. Белоусова. Второй выстрел]

(19) *Они были более сказочные.* — ***Вот именно***, — *пробубнил Лэри, вгрызаясь в бутерброд.* — *И я о том же толкую.* [Мариам Петросян. Дом, в котором…]

Коммуникативы *Именно* и *Именно, именно* также используются в аналогичных контекстах, являясь неполными реализациями фраземы *Вот именно1*. Ср.:

(20) — *Мы живём в век культуры,* — *ораторствовал Японец на другой день вечером.* — *Мы стремимся к прогрессу, а нам дают лапти. Презренные лапти!* — ***Именно!*** (= Вот именно) *Плевок прошлого,* — *мычал Пантелеев.* [Г. Г. Белых. Лапти]

(21) *Взять бы этого Канта, да за такие доказательства года на три в Соловки!* <…> — ***Именно, именно*** (= Вот именно!), — *закричал он, и левый зелёный глаз его, обращённый к Берлиозу, засверкал,* — *ему там самое место!* [М. А. Булгаков. Мастер и Маргарита]

В полемических диалогах отчетливо выявляется еще одно, риторическое употребление коммуникатива *вот именно2*. Собеседник, отстаивая свое мнение, приводит в его пользу **аргумент**. Говорящий этот аргумент горячо поддерживает, чтобы неожиданно для собеседника использовать его для продвижения собственного, часто альтернативного мнения.

(22) — *Как-то надо определиться.* <…> *Например, поступить на работу.* — *Зачем? Всё равно весной в армию.* — *Но сейчас ещё только осень.* — ***Вот именно***. *Осень, зиму погуляю. А там запрягут.* [И. Грекова. Перелом]

(23) — *Он тебе не нравится?* — *тихо спросила Снежана.* — *При чём тут я,* — *удивилась притворно Ирина.* — *Тебе жить.* —***Вот именно***, — *твёрдо сказала Снежана.* — *Я тебя очень прошу, не вмешивайся.* [Т. Виктория. Своя правда]

(24) — *Но, наверняка, он хочет жить дольше. — Кто ж не хочет — сказал я. —* ***Вот именно,*** *— засмеялся он радостно, — хотят все, да не всем дано.* [В. Войнович. Москва 2042]

Более того, говорящий может несколько нарушить правила коммуникативного взаимодействия, выбрав для мнимо солидарного *вот именно2* элемент из неассертивной части высказывания собеседника.

(25) — *Ты, з-з-зачем на мой цветок села? — опускаясь на душицу, прожужжала она. — Почему твой? Я его только что сама нашла. —* ***Вот именно*** *— только что. А я вчера…* [В. Кологрив. Медовый луг // «Мурзилка», 2002]

(26) *Мало того, что заставил меня через полмира переть, так еще тут выдрючивается. У него расписание, у него времени нет. Мне мое время в конце концов тоже для чего-то нужно. —* ***Вот именно*** *— оживился Зильберович. Твое время нужно тебе, а его время нужно всем, всему человечеству.* [В. Войнович. Москва 2042 год]

При формальной близости коммуникативов *Именно так* и *Вот именно* их взаимозамена приводит к прагматической ошибке, которую, видимо, стоит относить к просторечию. Такой тип ошибки может «царапать ухо» собеседнику, и иногда намеренно используется в художественной речи. Ср. диалог, на примере которого, в частности, можно говорить о своеобразии стиля Сергея Довлатова. При помощи одобрительного *Вот именно* иронический персонаж подтверждает (горячо одобряет) вопрос собеседника, как если бы это было оценочное суждение, утверждение, вывод.

(27) — *Решили поработать в заповеднике? —* ***Вот именно.*** [С. Довлатов. Заповедник]

Итак, поскольку, как было продемонстрировано в докладе, единое, предлагаемое в одном формате описание дискурсивов и коммуникативов приводит к затемнению полноценного описания коммуникативов, необходимо их отдельное лексикографическое описание, в котором будут отражены все условия использования этих диалогических единиц языка.

Литература

1. *АСРЯ 2014, т. 1* — Активный словарь русского языка. Т. 1. А—Б / Отв. ред. акад. Ю. Д. Апресян. М.: Языки славянской культуры, 2014.
2. *Баранов, Добровольский 2009* — Фразеологический объяснительный словарь русского языка. Под. ред. А. Н. Баранова, Д. О. Добровольского. М., 2009.

3. *Баранов, Кобозева 1988* — Баранов А. Н., Кобозева И. М. Модальные частицы в ответах на вопрос // Прагматика и проблемы интенсиональности. М. : ИВАН СССР, 1988. С. 45–69.
4. *Галактионова 1995* — Галактионова И. В. Средства выражения согласия/несогласия в русском языке: АКД. М., 1995. 18 с.
5. *Галактионова 1995* — Галактионова И. В. Средства выражения согласия/несогласия в русском языке: АКД. М., 1995. 18 с.
6. *Добрушина 1993* — Добрушина Е. Р. Семантика частиц ну и ну да для обозначения верификативности в разговорной речи // Семантика языковых единиц. Материалы 3-ей межвузовской научно-исследовательской конференции. Часть III Синтаксическая семантика. М., 1993.
7. *Добрушина 1993* — Добрушина Е. Р. Верификация в современной диалогической речи. АКД. М., 1993. 20 с.
8. *ДСРЯ 1998* — Дискурсивные слова русского языка: опыт контекстно-семантического описания. Под ред. К. Киселевой и Д. Пайара. М., 1998.
9. *Квеселевич, Сасина 1990* — Д. И. Квеселевич, В. П. Сасина. Русско-английский словарь междометий и релятивов. М., 1990.
10. *Колокольцева 2001* — Колокольцева Т. Н. Специфические коммуникативные единицы диалогической речи. Волгоград 2001.
11. *Кустова 2012* — Кустова Г. И. Об иллокутивной фразеологии // Смыслы, тексты и другие захватывающие сюжеты. Сборник статей в честь 80-летия И. А. Мельчука. М., «Языки славянской культуры», 2012, с. 349–365.
12. *Михайлова 1999* — Михайлова Е. А. Многокомпонентные реплики-частицы, выражающие согласие, несогласие и верификацию в русской диалогической речи. АКД. Казань. 1999
13. *Молотков 1986* — Фразеологический словарь русского языка. Под ред. А. И. Молоткова. 4-е изд. М., 1986.
14. *НОССРЯ 1997, 2000, 20003* — Апресян, Ю. Д. и др. 2000. Новый объяснительный словарь синонимов русского языка. Под общим руководством Ю. Д. Апресяна. Выпуски 1, 2, 3. Москва: Языки русской культуры.
15. *Ожегов 1990* — Ожегов С. И. Словарь русского языка / под ред. Н. Ю. Шведовой., М.: Рус. яз., 1990.
16. *Падучева 1982* — Падучева Е. В. Прагматические аспекты связности диалога. СЛЯ, том 41, № 4. 1982.
17. *ПДС 1993* — Баранов А. Н., Плунгян В. А., Рахилина Е. В. Путеводитель по дискурсивным словам русского языка. М., 1993.
18. *Разлогова 1996* — Разлогова Е. Модальные слова и оценка степени достоверности высказывания// Русистика сегодня №3/96.
19. *Рогожникова 1983* — Рогожникова Р. П. Словарь сочетаний, эквивалентных слову. М., 1983.
20. *СССРЯ 1997* — Морковкин В. В., Н. М. Луцкая, Г. Ф. Богачева. Словарь структурных слов русского языка. / Под ред. В. В. Морковкина. М., 1997.
21. *Шаронов 1996* — Шаронов И. А. Коммуникативы как функциональный класс и объект лексикографического описания// Русистика сегодня №2, 1996, с. 89–112.

22. *Шаронов 2015* — Шаронов И. А. Поиск и описание коммуникативов на основе национального корпуса русского языка // «Методы когнитивного анализа семантики слова: компьютерно-корпусный подход». Коллективная монография. Изд. «Языки славянских культур», 2015. Отв. ред. В. И. Заботкина. Гл. 5, с. 145–187.
23. *Шимчук, Щур 1999* — Шимчук. Э., Щур М. Словарь русских частиц / Berliner slavistishe Arbeiten. B. 9. Frankfurt am Main, 1999.
24. *Яковлева 1994* — Е. С. Яковлева. Фрагменты русской языковой картины мира. М., 1994.
25. *Lubensky 1995* — S. Lubensky. Russian-English Dictionary of Idioms. Random house. NY. 1995.
26. *Paukkeri 2006* — Paukkeri P. Реципиент в русском разговоре: о распределении функций: между ответами да, ну и так. Esitetään Helsingin yliopiston humanistisen tiedekunnan suostumuksella julkisesti tarkastettavaksi Metsätalon salissa 1 perjantaina 26. toukokuuta 2006 klo 12. Helsinki 2006. Slavica Helsingensia 28.
27. *Sorjonen 2001* — Sorjonen, M.-L. 2001: Responding in Conversation. A study of response particles in Finnish. Amsterdam/Philadelphia: John Benjamins.
28. *Sorjonen 2002* — Recipient activities: the particle no as a go-ahead response in Finnish conversations. The language of turn and sequence. C. E. Ford, B. A. Fox & S. A.

НАИБОЛЕЕ УПОТРЕБИТЕЛЬНЫЕ СЛОВА ПОВСЕДНЕВНОЙ РУССКОЙ РЕЧИ (В ГЕНДЕРНОМ АСПЕКТЕ И В ЗАВИСИМОСТИ ОТ УСЛОВИЙ КОММУНИКАЦИИ)[1]

Шерстинова Т. Ю. (t.sherstinova@spbu.ru)

СПбГУ, Санкт-Петербург, Россия

В центре внимания данной работы находятся наиболее употребительные слова русской повседневной речи, представляющие собой верхнюю зону частотных словарей, полученных на материале звукозаписей речевого корпуса «Один речевой день» (ОРД). Речевой материал, представленный в корпусе, проаннотирован с точки зрения условий коммуникации (тип коммуникации/языковой стиль, социальная роль говорящего, локус и др.), а также снабжен информацией о социальных характеристиках информанта и его основных коммуникантов. Такая информация позволяет фильтровать речевой материал и исследовать изменение речевых характеристик в зависимости от социальных характеристик говорящего и условий коммуникации. Исследование выполнено на материале 152 эпизодов повседневной речевой коммуникации, которые содержат в общей сложности 232 370 словоупотреблений. Выборка содержит речь 209 говорящих (95 мужчин, 94 женщины, 20 детей). Построены общий частотный словарь, мужской и женские словари устной речи, а также гендерные словари для четырех стилей устной речи: 1) бытовые неформальные разговоры, 2) профессиональные (деловые/официальные) разговоры, 3) учебно-образовательная речевая коммуникация, 4) коммуникация по типу «клиент-сервис».

Ключевые слова: русская разговорная речь, повседневная речевая коммуникация, частотные словари, социолингвистика, гендерная вариативность речи, стили устной речи, речевой корпус, устный дискурс

THE MOST FREQUENT WORDS IN EVERYDAY SPOKEN RUSSIAN (IN THE GENDER DIMENTION AND DEPENDING ON COMMUNICATION SETTINGS)

Sherstinova T. Yu. (t.sherstinova@spbu.ru)

Saint-Petersburg State University, Saint-Petersburg, Russia

[1] Исследование выполнено при поддержке гранта РНФ № 14-18-02070 «Русский язык повседневного общения: особенности функционирования в разных социальных группах».

The paper presents the most frequent words of everyday spoken Russian, that form the upper zones of several word frequency lists compiled on the material of Russian speech corpus “One Speaker’s Day” (the ORD corpus), containing real-life recordings of everyday communication. All speech data in the corpus is annotated in terms of communication settings, including 1) type of communication (language spoken style), 2) social role of speaker, 3) locus, etc. Such information allows speech to be filtered upon user request and therefore makes it possible to study speech variation depending on particular communication settings. The given study was made on the transcripts of 152 real-life macroepisodes and contains 232,370 words. The sample presents speech of 209 persons (95 men, 94 women, 20 children). The following word frequency lists have been compiled: a) general frequency list, b) male frequency list, c) female frequency list, and d) four frequency lists for different styles of spoken speech: informal conversations, professional/business conversations, educational communication, and “customer-service” communication. Men’s and women’s frequency lists have been compiled on the subsamples of 83,371 and 115,110 words correspondently. The analysis of word lists has shown that Russian women pay more attention to maintaining the conversation, use fewer hesitations, and are more inclined to use in their speech intensifying words, emotional words, hedges and interjections. Men generally use fewer personal pronouns, while numbers and the expletives are among the most frequent words used by men in everyday conversations. In general, these observations are similar to those described earlier for gender variation by other linguists.

Key words: spoken Russian, everyday verbal communication, word frequency lists, sociolinguistics, gender variability of speech, styles of spoken language, speech corpus, oral discourse

1. Введение

Построение частотных словарей — традиционный метод современных лексикографических исследований на базе лингвистических корпусов (см., например, [Мартыненко 1988; Leech et. al 2001; Popescu 2009; Ляшевская, Шаров, 2009; Шайкевич 2015] и др.). Частоту языковых единиц можно рассматривать как индикатор маркированности [Baker, 2010, p. 125], который позволяет оценить их функциональную активность. Данные, представленные в «сухой» табличной форме частотного словаря, лаконично и по существу выявляют особенности текстового материала и позволяют проводить его статистический анализ. Особый интерес исследователей привлекает «верхняя зона» частотных словарей, представляющая костяк речевой системы, во многом совпадающая не только для индивидуальных говорящих, но и для целых подсистем (языковых стилей, регистров, ситуаций общения). Для выявления таких зон предлагаются специальные индексы [Мартыненко 1988, Popescu 2009].

За последние десятилетия подготовлено большое количество частотных словарей как для отдельных языковых жанров (научной и художественной литературы, специальных текстов), так и для отдельных выдающихся писателей. Однако до сих пор лингвисты не располагают статистически представительными данными о частоте слов в наиболее важном для человека языковом жанре — живой

спонтанной речи, которая составляет основу повседневной коммуникации. Предлагаемое исследование позволяет отчасти восполнить этот пробел, представляя частотные списки словоформ в зависимости от разных стилей повседневной устной речи в речи мужчин и женщин, в разных ситуациях общения.

Другой целью данной работы является демонстрация возможностей речевого корпуса ОРД, предоставляющего возможность выборки и анализа данных не только по социодемографическим характеристикам говорящего и стилю устной речи, являющихся традиционными для многих речевых корпусов, но также и в зависимости от разных условий коммуникации [Sherstinova 2015].

2. Речевой корпус ОРД

Разработка речевого корпуса повседневной русской речи ОРД («Один речевой день»), для получения звукозаписей которого информанты-добровольцы согласились прожить целый день с «диктофоном на шее», записывающим всю их речевую коммуникацию, позволяет вывести исследования повседневной устной речи на качественно новый уровень [Asinovsky et al. 2009]. Подобная методика записи речи традиционно используется в японских лингвистических исследованиях (см., например, [Campbell 2004]), а также применялась при подготовке материалов для устного подкорпуса BNC [Burnard 2007]. Ее преимущество состоит в получении для анализа речевого материала, наиболее приближенного к естественной повседневной речи.

В настоящее время корпус содержит более 1200 часов звукозаписей речи, полученной от 127 информантов, мужчин и женщин, в возрасте от 17 до 77 лет, и нескольких сотен их коммуникантов. Сбор материала и аннотирование корпуса продолжается [Bodganova-Beglarian et al. 2015].

Все звукозаписи «речевого дня» подлежат сегментации на макроэпизоды повседневного общения, которые аннотируются с точки зрения условий коммуникации (локуса, социальных ролей участников, типа коммуникации и некоторых ее особенностей) [Sherstinova 2015]. Нормализация соответствующих кодов позволяет использовать эти параметры в качестве фильтров для выборки необходимого речевого материала. Благодаря этим фильтрам были получены представленные в работе списки частотных слов.

3. Материал и методика

Данное разведочное исследование выполнено на расшифровках 152 макроэпизодов, записанных 40 информантами и их коммуникантами. Записи выполнены в Санкт-Петербурге в 2007 и 2010 гг. Выборка содержит речь 209 человек:

Информанты		**Основные коммуниканты**				**Итого (человек)**
Муж.	**Жен.**	**Муж.**	**Жен.**	**Дети**	**Всего**	
20	20	75	74	20	169	209

Общая длительность звучания проанализированного речевого материала составляет 40,5 часов, в среднем по 3,8 эпизода и 1 часу звукозаписи от каждого информанта. Объем подкорпуса в словоупотреблениях составляет 232 370 единиц. Средняя продолжительность эпизода — 15,99 мин. (SD = 9,80 мин.), в словах — 1550 (SD = 1120).

По типам коммуникации, коррелирующим с соответствующими стилями устной речи, эпизоды исследуемой выборки имеют следующее распределение:

Тип коммуникации / Стиль устной речи	**Процент эпизодов**	**Кол-во эпизодов**	**Кол-во словоупотр.**
Бытовые неформальные разговоры	62,5%	95	154 051
Профессиональные (деловые) разговоры	19,7%	30	40 012
Коммуникация по типу «клиент-сервис»	9,9%	15	21 126
Учебная коммуникация	7,2%	11	19 629
Публичные выступления	0,7%	1	800

Учебная коммуникация понимается как общение между «обучающим» и «обучаемым» (лекции, практические занятия, индивидуальные занятия, инструктаж, обучающее занятие с ребенком и т. п.). Коммуникация по типу «клиент-сервис», как правило, представляет собой относительно формальный разговор между человеком, профессионально оказывающим «услуги» в широком смысле слова (кроме обучающих) и его клиентом (в государственных службах, сервис-центрах, магазинах, поликлиниках, библиотеках и т. п.). Профессиональная беседа — деловой разговор на профессиональные темы, не являющийся ни учебной, ни «клиент-сервисной» коммуникацией, осуществляемый преимущественно между коллегами. Наконец, «бытовой разговор» имеет неформальный характер, в большинстве случаев не связан тематически с профессиональной деятельностью информанта и может иметь место среди разных коммуникантов, в разных условиях общения. Как любое прагматическое аннотирование, атрибуция макроэпизодов по типам коммуникации в большой степени зависит от контекста и содержания беседы.

Полученное распределение в целом согласуется с распределением эпизодов по типу коммуникации, посчитанных для корпуса ОРД на материале 1854 макроэпизодов, описывающих 483 часов звучащей речи [Sherstinova 2015].

Следует отметить следующие особенности построения частотных списков:

1) В связи с тем, что процесс лемматизации и разведения омонимии корпуса ОРД еще не завершен, здесь анализируются словоформы, без учета омонимичных форм. Впрочем, словоформы довольно часто используются при анализе частот [Rayson et al. 1997; Popescu 2009]. В нашем случае «смягчающим» обстоятельством, оправдывающим такой подход, является тот факт, что рассматриваемая нами верхняя зона частотного словаря состоит преимущественно из неизменяемых частиц. Однако очевидно, что при лемматизации следует ожидать роста частот личных местоимений — другой высокоактивной категории единиц.

2) Все слова в транскриптах ОРД записаны в стандартной орфографии, независимо от реального произнесения слова («*сейчас*», а не «*щас*»; «*что*», а не «*чё*» или «*шо*»), поскольку особенности фонетической реализации слов в корпусе ОРД отмечаются на специальных фонетических уровнях.
3) В представленных частотных списках было решено оставить и некоторые «несловарные» элементы устной речи (в частности, «*угу*» и «*ага*», а также заполненные (*э*) и незаполненные (...) паузы хезитации).
4) При интерпретации данных следует иметь в виду, что двухсловные языковые единицы, такие как «*потому что*», «*так как*» и т. п., как и в большинстве других частотных словарей, представлены здесь отдельными «частями».
5) В качестве «наиболее употребительных» слов рассматривается первая сотня словоформ верхней зоны соответствующего частотного словаря.

4. Общий частотный словарь

Прежде всего рассмотрим распределение наиболее активной лексики в целом по выборке при отсутствии каких-либо фильтров.

В табл. 1 каждое слово сопровождается численной информацией: ранг слова, его абсолютная частота по выборке, доля (в процентах), а также кумулятивный процент. Последний показатель довольно интересен, так как позволяет оценить совокупную долю частот с рангом не ниже данной. Так, из табл. 1 видно, что 10 наиболее употребительных слов устной речи (*я*, *вот*, *ну*, *не*, *да*, *а*, *и*, *что*, *в*, *это*) покрывают примерно 1/5 (20,69 %) всей речевой коммуникации, а 88 самых частотных слов покрывают уже половину (50,00 %) всего речевого материала.

Наиболее употребительными словами устной русской речи на нашем материале оказались личные местоимения (*я*, *ты*, *он*, *она*, *они*, *мы*, *вы* в форме им. п., *мне* (дат./предл.), *меня* (род./вин. п.)) и неизменяемые единицы — частицы, союзы, предлоги и слова, которые в последнее время все чаще стали называть «дискурсивными словами» или «маркерами» (*вот*, *ну*, *не*, *да*, *нет*, *так*, *просто, вообще*, *значит*, *сейчас*, и др.). Многие из этих единиц выполняют в речи прагматические функции и/или используются для регулирования дискурса. Среди частотных мы видим и заполненную хезитацию (*э*). Другая часть высокочастотных слов репрезентирует кластеры омонимичных форм, аттрибутировать которые невозможно без обращения к контексту (*а*, *есть*, *что* и др.).

Необходимо отметить, что полученные данные в определенной степени коррелируют с другими частотными словарями устной речи, подготовленными на материале русского [Ляшевская, Шаров 2009] и английского языков [Leech et al. 2001], но и имеют свою специфику.

Таблица 1. Наиболее употребительные слова устной русской речи (общий список)

Ранг	Слово	Абс. ч.	%	Кумул. %	Ранг	Слово	Абс. ч.	%	Кумул. %
1	я	5950	2,56	2,56	51	очень	553	0,24	42,97
2	вот	5411	2,33	4,89	52	все	550	0,24	43,21
3	ну	5407	2,33	7,22	53,5	потом	545	0,23	43,45
4	не	5400	2,32	9,54	53,5	тебе	545	0,23	43,68
5	да	5162	2,22	11,76	55	говорит	542	0,23	43,91
6	а	4523	1,95	13,71	56	может	536	0,23	44,14
7	и	4190	1,80	15,51	57	хорошо	532	0,23	44,37
8	что	4139	1,78	17,29	58	когда	527	0,23	44,60
9	в	4003	1,72	19,01	59	или	525	0,23	44,83
10	это	3905	1,68	20,69	60	его	496	0,21	45,04
11	там	3422	1,47	22,17	61	такой	489	0,21	45,25
12	у	3007	1,29	23,46	62	тебя	485	0,21	45,46
13	так	2842	1,22	24,68	63	за	480	0,21	45,66
14	на	2602	1,12	25,80	64	давай	468	0,20	45,87
15	как	2110	0,91	26,71	65	говорю	465	0,20	46,07
16	ты	1877	0,81	27,52	66	только	464	0,20	46,27
17	всё	1874	0,81	28,33	67	ой	463	0,20	46,47
18	то	1781	0,77	29,09	68	можно	450	0,19	46,66
19	с	1771	0,76	29,85	69,5	потому	449	0,19	46,85
20	нет	1727	0,74	30,60	69,5	знаешь	449	0,19	47,05
21	(э)	1708	0,74	31,33	71	где	432	0,19	47,23
22	он	1670	0,72	32,05	72	бл**ь	430	0,19	47,42
23	угу	1451	0,62	32,68	73	ага	426	0,18	47,60
24	мне	1293	0,56	33,23	74	ничего	425	0,18	47,78
25	она	1275	0,55	33,78	75,5	конечно	405	0,17	47,96
26	есть	1256	0,54	34,32	75,5	что-то	405	0,17	48,13
27	меня	1124	0,48	34,81	77	этот	401	0,17	48,30
28	сейчас	1117	0,48	35,29	78	чего	390	0,17	48,47
29	они	1060	0,46	35,74	79	вас	377	0,16	48,63
30	мы	1050	0,45	36,19	80	такая	369	0,16	48,79
31	бы	1002	0,43	36,63	81	раз	366	0,16	48,95
32	но	989	0,43	37,05	82	такое	359	0,15	49,10
33	уже	984	0,42	37,47	83	до	357	0,15	49,26
34	надо	953	0,41	37,88	84	её	356	0,15	49,41
35	ещё	938	0,40	38,29	85	чтобы	343	0,15	49,56
36	же	914	0,39	38,68	86	вам	342	0,15	49,71
37	по	889	0,38	39,06	87	эти	341	0,15	49,85
38	просто	875	0,38	39,44	88	даже	340	0,15	50,00
39	вообще	772	0,33	39,77	89	их	334	0,14	50,14
40	если	721	0,31	40,08	90	блин	328	0,14	50,28
41	вы	712	0,31	40,39	91	был	321	0,14	50,42
42	нас	691	0,30	40,69	92	два	312	0,13	50,56

Ранг	Слово	Абс. ч.	%	Кумул. %	Ранг	Слово	Абс. ч.	%	Кумул. %
43	тоже	665	0,29	40,97	93	для	309	0,13	50,69
44	знаю	664	0,29	41,26	94	один	305	0,13	50,82
45	было	592	0,25	41,51	95	кто	303	0,13	50,95
46	значит	584	0,25	41,76	96	быть	300	0,13	51,08
47	здесь	577	0,25	42,01	97,5	из	299	0,13	51,21
48	будет	570	0,25	42,26	97,5	ладно	299	0,13	51,34
49	к	556	0,24	42,50	99	ли	297	0,13	51,47
50	тут	555	0,24	42,74	100	короче	296	0,13	51,59

5. Наиболее частотные слова мужской и женской речи

Исследование гендерной вариативности речи представляет собой популярное направление в российской и мировой лингвистике [Lakoff, 1975; Tannen, 1991; Потапова, Потапов 2006; и мн. др.]). Ее изучение на материале ОРД возможно благодаря сбалансированности информантов корпуса по гендерному параметру.

Так, мужская и женская речь представлены в анализируемой выборке практически одинаковым количеством говорящих, однако объем «женского подкорпуса» в словах превысил «мужской» примерно на треть, что дает основания предполагать, что мужчины в среднем говорят меньше. С другой стороны, индекс лексического богатства для речи мужчин оказался несколько выше, чем у женщин:

	Количество говорящих	Всего словоупотреблений[2]	Всего разных словоформ	Всего однократных форм	Индекс лексического богатства
Мужской подкорпус	95	83 371	14 539	9 408	0,174
Женский подкорпус	94	115 110	17 470	10 927	0,152

В табл. 2 и 3 приводятся наиболее употребительные слова женской и мужской устной русской речи соответственно.

Первое, что бросается в глаза, это отличие слов «первого» ранга. Личное местоимение *я*, являющееся абсолютным лидером в женской речи, уступает пальму первенства частице *ну*[3] в мужской речи. Видно, что женщины уделяют

---

[2] В гендерные подкорпуса по сравнению с общей выборкой не была включена речь детей и подростков до 18 лет, а также ряд фрагментов, характеризующиеся одновременной речью двух или более говорящих.

[3] Проведенное ранее исследование позволяет предположить, что использование частицы «*ну*» более характерно для неформального общения, в отличие от частицы «*вот*», которая чаще используется в официальной (деловой) речи (Sherstinova 2015), однако эта гипотеза требует проверки.

больше внимания формальному поддержанию разговора (*угу*, *хорошо*), меньше хезитируют и чаще используют усилительные слова (*очень*) и междометия (*ой*).

Мужчины реже употребляют в речи личные местоимения, у них чаще наблюдаются хезитации. Наиболее отличительной чертой мужской речи по сравнению с женской является отмечаемое всеми без исключениями исследователями [Lakoff 1975, Stenström 1991, Rayson et al. 1997, и др.] появление в верхней зоне частотного словаря бранной лексики, непечатных слов и их субститутов.

В связи с тем, что мужской и женский словари не совпадают по составу лексических единиц, в том числе и в «верхней зоне», сравнение частоты употребления проводится относительно или первого, или второго списка. Так, например, выглядят пятерки «преимущественно женских» и «мужских» слов:

«Наиболее женские» слова				**«Наиболее мужские» слова**			
По разности рангов		**По разности долей**		**По разности рангов**		**По разности долей**	
нам	68,5	угу	0,377%	х**	–	б**дь	0,504%
как-то	61,5	так	0,370%	б**дь	3365	в	0,464%
слушай	60,5	что	0,326%	блин	283	там	0,452%
вам	49	да	0,322%	короче	98,5	на	0,313%
ой	40	я	0,303%	типа	72,5	блин	0,311%

В целом наблюдаемые на материале ОРД различия в речи мужчин и женщин совпадают с выводами, сделанными ранее, в том числе и на материале других языков [Coupland 2007; Romaine 2008; Потапова, Потапов 2006; и др.].

Таблица 2. Наиболее употребительные слова женской устной русской речи

Ранг	**Слово**	**Абс. ч.**	**%**	**Кумул. %**	**Ранг**	**Слово**	**Абс. ч.**	**%**	**Кумул. %**
1	я	3110	2,70	2,70	51,5	знаешь	274	0,24	43,59
2	не	2722	2,36	5,07	51,5	все	274	0,24	43,83
3	вот	2691	2,34	7,40	53	говорит	271	0,24	44,06
4	да	2593	2,25	9,66	54	потом	268	0,23	44,29
5	ну	2577	2,24	11,90	55	тебе	262	0,23	44,52
6	что	2218	1,93	13,82	56	или	260	0,23	44,75
7	а	2175	1,89	15,71	57	может	259	0,23	44,97
8	и	2116	1,84	17,55	58	давай	255	0,22	45,19
9	это	1925	1,67	19,22	59	говорю	254	0,22	45,41
10	в	1770	1,54	20,76	60	здесь	246	0,21	45,63
11	так	1619	1,41	22,17	61	тут	245	0,21	45,84
12	у	1569	1,36	23,53	62	когда	241	0,21	46,05
13	там	1490	1,29	24,82	63	только	238	0,21	46,26
14	на	1176	1,02	25,85	64,5	его	236	0,21	46,46
15	как	1140	0,99	26,84	64,5	ничего	236	0,21	46,67
16	ты	1004	0,87	27,71	66	что-то	235	0,20	46,87

Ранг	Слово	Абс. ч.	%	Кумул. %	Ранг	Слово	Абс. ч.	%	Кумул. %
17	всё	932	0,81	28,52	67	такой	231	0,20	47,07
18	с	913	0,79	29,31	68	тебя	230	0,20	47,27
19	угу	836	0,73	30,04	69	где	223	0,19	47,47
20	то	826	0,72	30,75	70,5	значит	220	0,19	47,66
21	нет	798	0,69	31,45	70,5	конечно	220	0,19	47,85
22	она	783	0,68	32,13	72	вас	218	0,19	48,04
23	мне	749	0,65	32,78	73	потому	217	0,19	48,23
24	он	739	0,64	33,42	74	можно	214	0,19	48,41
25	(э)	677	0,59	34,01	75,5	за	208	0,18	48,59
26	мы	651	0,57	34,57	75,5	ага	208	0,18	48,77
27	сейчас	571	0,50	35,07	77	такая	198	0,17	48,95
28	меня	568	0,49	35,56	78	вам	197	0,17	49,12
29	есть	564	0,49	36,05	79	её	195	0,17	49,29
30	бы	561	0,49	36,54	80	такое	194	0,17	49,45
31	но	523	0,45	37,00	81	даже	171	0,15	49,60
32	они	518	0,45	37,45	82	ладно	167	0,15	49,75
33	надо	486	0,42	37,87	83,5	чтобы	166	0,14	49,89
34	ещё	477	0,41	38,28	83,5	чего	166	0,14	50,04
35	уже	469	0,41	38,69	85,5	этот	164	0,14	50,18
36	же	452	0,39	39,08	85,5	эти	164	0,14	50,32
37	по	430	0,37	39,46	87	их	162	0,14	50,46
38,5	вообще	423	0,37	39,82	88,5	нам	161	0,14	50,60
38,5	вы	423	0,37	40,19	88,5	сегодня	161	0,14	50,74
40	просто	407	0,35	40,54	90,5	раз	159	0,14	50,88
41	нас	395	0,34	40,89	90,5	до	159	0,14	51,02
42	тоже	364	0,32	41,20	92	как-то	158	0,14	51,16
43	знаю	352	0,31	41,51	93	для	157	0,14	51,29
44	очень	342	0,30	41,81	94	ли	152	0,13	51,42
45	если	326	0,28	42,09	95	о	148	0,13	51,55
46	к	298	0,26	42,35	96	них	147	0,13	51,68
47	будет	294	0,26	42,60	97	слушай	145	0,13	51,81
48	хорошо	292	0,25	42,86	98	пока	144	0,13	51,93
49	было	288	0,25	43,11	99,5	быть	143	0,12	52,06
50	ой	278	0,24	43,35	99,5	думаю	143	0,12	52,18

Таблица 3. Наиболее употребительные слова мужской устной русской речи

Ранг	Слово	Абс. ч.	%	Кумул. %	Ранг	Слово	Абс. ч.	%	Кумул. %
1	ну	2062	2,47	2,47	50,5	было	208	0,25	42,29
2	я	2000	2,40	4,87	52	потом	207	0,25	42,54
3	не	1898	2,28	7,15	53	все	205	0,25	42,78
4	вот	1887	2,26	9,41	54,5	короче	201	0,24	43,02
5	в	1669	2,00	11,41	54,5	тоже	201	0,24	43,26

Ранг	Слово	Абс. ч.	%	Кумул. %	Ранг	Слово	Абс. ч.	%	Кумул. %
6	да	1610	1,93	13,35	56	такой	194	0,23	43,50
7	а	1548	1,86	15,20	57	тебя	192	0,23	43,73
8	и	1530	1,84	17,04	58	будет	191	0,23	43,96
9	там	1456	1,75	18,78	59	нас	189	0,23	44,18
10	это	1338	1,60	20,39	60	или	183	0,22	44,40
11	что	1335	1,60	21,99	61	чего	182	0,22	44,62
12	на	1113	1,33	23,32	62	его	177	0,21	44,83
13	у	985	1,18	24,51	63	может	173	0,21	45,04
14	так	864	1,04	25,54	64	только	172	0,21	45,25
15	(э)	761	0,91	26,46	65	этот	169	0,20	45,45
16	он	696	0,83	27,29	66	потому	166	0,20	45,65
17	то	680	0,82	28,11	67	к	165	0,20	45,85
18	всё	665	0,80	28,90	68	вы	157	0,19	46,04
19	как	656	0,79	29,69	69	хорошо	156	0,19	46,22
20	ты	651	0,78	30,47	70	давай	153	0,18	46,41
21	с	609	0,73	31,20	71	раз	147	0,18	46,58
22	нет	537	0,64	31,85	72,5	один	146	0,18	46,76
23	есть	478	0,57	32,42	72,5	говорю	146	0,18	46,93
24	бл**ь	422	0,51	32,93	74	где	144	0,17	47,11
25	сейчас	410	0,49	33,42	75	можно	143	0,17	47,28
26	меня	393	0,47	33,89	76	кто	139	0,17	47,44
27	они	387	0,46	34,35	77,5	был	137	0,16	47,61
28	мне	356	0,43	34,78	77,5	знаешь	137	0,16	47,77
29	уже	344	0,41	35,19	79,5	ничего	134	0,16	47,93
30	просто	331	0,40	35,59	79,5	их	134	0,16	48,09
31	ещё	330	0,40	35,98	81	очень	133	0,16	48,25
32	но	325	0,39	36,37	82	два	132	0,16	48,41
33	надо	320	0,38	36,76	83	до	131	0,16	48,57
34	же	319	0,38	37,14	84,5	чтобы	127	0,15	48,72
35	она	318	0,38	37,52	84,5	такая	127	0,15	48,87
36	по	315	0,38	37,90	86	типа	126	0,15	49,02
37	значит	294	0,35	38,25	87	такое	125	0,15	49,17
38	угу	291	0,35	38,60	88,5	ага	124	0,15	49,32
39	блин	283	0,34	38,94	88,5	из	124	0,15	49,47
40	если	281	0,34	39,28	90	ой	122	0,15	49,62
41	бы	274	0,33	39,61	91	даже	121	0,15	49,76
42	вообще	269	0,32	39,93	92,5	что-то	119	0,14	49,91
43	знаю	248	0,30	40,23	92,5	конечно	119	0,14	50,05
44	мы	228	0,27	40,50	94,5	её	118	0,14	50,19
45,5	тут	218	0,26	40,76	94,5	х**	118	0,14	50,33
45,5	говорит	218	0,26	41,02	96	три	117	0,14	50,47
47	здесь	215	0,26	41,28	97	эти	115	0,14	50,61
48	за	213	0,26	41,54	98	для	110	0,13	50,74
49	когда	210	0,25	41,79	99	понятно	108	0,13	50,87
50,5	тебе	208	0,25	42,04	100	тогда	107	0,13	51,00

В качестве примера того, как может перераспределяться функциональная активность слов в зависимости от коммуникативной ситуации, приведем 2 таблицы, в которых показаны по 50 наиболее частотных слов тех же мужского и женского подкорпусов для разных языковых стилей[4]: 1) неформальной бытовой речи, 2) профессионального разговора, 3) «учебной» коммуникации и 4) широкого пласта коммуникативных ситуаций по типу «клиент-сервис».

Поскольку при дроблении выборки на подкатегории статистическая представительность каждой из них неизбежно уменьшается, представленные данные носят скорее иллюстративный характер (см. табл. 4 и 5).

	Объем подкорпусов разных стилей (словоупотребления)			
	Бытовой	**Профессиональный**	**«Клиент-сервис»**	**Учебный**
Женский подкорпус	72 759	25 239	6 838	10 094
Мужской подкорпус	62 030	10 123	5 187	5 254

Таблица 4. Наиболее употребительные слова женской устной речи в зависимости от коммуникативной ситуации (стиля речи)

Бытовая			**Профессиональная**			**Образовательная**			**«Клиент-сервис»**		
Ранг	**Слово**	**%**	**Ранг**	**Слово**	**%**	**Ранг**	**Слово**	**%**	**Ранг**	**Слово**	**%**
1	я	2,87	1	я	2,92	1	да	3,35	1	вот	3,01
2	не	2,61	2	вот	2,59	2	вот	3,05	2	я	2,28
3	ну	2,46	3	что	2,27	3	ну	1,85	3	не	2,19
4	да	2,23	4	не	2,21	4	так	1,80	4	и	1,97
5	а	2,17	5	и	2,11	5	это	1,63	5	да	1,97
6	вот	2,09	6	да	1,96	6	что	1,60	6	это	1,95
7	что	1,88	7	ну	1,87	7	у	1,43	7	а	1,87
8	и	1,81	8	в	1,75	8	и	1,31	8	ну	1,81
9	это	1,66	9	это	1,65	9	я	1,24	9	что	1,59
10	в	1,57	10	а	1,44	10	не	1,14	10	угу	1,52
11	там	1,40	11	там	1,34	11	мы	1,06	11	в	1,43
12	так	1,40	12	так	1,30	12	вы	1,01	12	у	1,36
13	у	1,40	13	у	1,23	13	угу	1,01	13	так	1,29
14	на	1,16	14	(э)	1,13	14	а	1,00	14	всё	1,10
15	ты	1,08	15	то	1,05	15	всё	0,92	15	как	1,10
16	как	1,02	16	всё	1,01	16	там	0,89	16	то	0,98
17	с	0,83	17	как	0,92	17	по	0,89	17	есть	0,94
18	он	0,74	18	она	0,90	18	как	0,86	18	на	0,88
19	нет	0,73	19	на	0,82	19	в	0,84	19	вам	0,85
20	всё	0,70	20	с	0,78	20	(э)	0,70	20	с	0,78

[4] Подробнее о стилях см. п. 3.

Бытовая			Профессиональная			Образовательная			«Клиент-сервис»		
Ранг	**Слово**	**%**	**Ранг**	**Слово**	**%**	**Ранг**	**Слово**	**%**	**Ранг**	**Слово**	**%**
21	мне	0,68	21	есть	0,74	21	на	0,65	21	она	0,76
22	она	0,67	22	мне	0,73	22	с	0,57	22	вы	0,75
23	угу	0,63	23	угу	0,66	23	нас	0,56	23	нет	0,73
24	то	0,63	24	нет	0,65	24	сейчас	0,51	24	сейчас	0,72
25	мы	0,57	25	ты	0,57	25	вас	0,50	25	бы	0,67
26	они	0,53	26	меня	0,56	26	давай	0,49	26	мне	0,63
27	бы	0,53	27	сейчас	0,53	27	нет	0,47	27	(э)	0,60
28	меня	0,52	28	вы	0,53	28	здесь	0,47	28	мы	0,58
29	но	0,48	29	но	0,47	29	он	0,44	29	он	0,57
30	сейчас	0,46	30	он	0,46	30	можно	0,42	30	вас	0,57
31	вообще	0,45	31	бы	0,46	31	вам	0,40	31	там	0,56
32	ещё	0,44	32	по	0,45	32	если	0,40	32	же	0,53
33	надо	0,44	33	надо	0,44	33	есть	0,39	33	уже	0,53
34	же	0,43	34	очень	0,43	34	то	0,39	34	надо	0,51
35	уже	0,42	35	просто	0,40	35	к	0,39	35	ты	0,51
36	есть	0,38	36	ещё	0,40	36	ты	0,39	36	можно	0,48
37	(э)	0,38	37	уже	0,38	37	или	0,38	37	но	0,47
38	тоже	0,35	38	они	0,36	38	может	0,36	38	ещё	0,44
39	нас	0,35	39	хорошо	0,35	39	будет	0,34	39	по	0,42
40	знаю	0,34	40	мы	0,34	40	дальше	0,32	40	меня	0,41
41	просто	0,34	41	будет	0,32	41	просто	0,32	41	если	0,39
42	ой	0,30	42	когда	0,31	42	хорошо	0,31	42	просто	0,38
43	тебе	0,28	43	знаю	0,31	43	уже	0,29	43	тоже	0,35
44	было	0,28	44	значит	0,30	44	потом	0,29	44	они	0,35
45	говорит	0,28	45	же	0,30	45	же	0,28	45	только	0,34
46	знаешь	0,27	46	вообще	0,29	46	(...)	0,27	46	говорю	0,32
47	очень	0,27	47	потому	0,29	47	но	0,26	47	нас	0,31
48	по	0,27	48	если	0,28	48	смотрим	0,26	48	потому	0,31
49	если	0,26	49	может	0,26	49	уровень	0,25	49	будет	0,29
50	к	0,25	50	все	0,26	50	значит	0,25	50	хорошо	0,29

Таблица 5. Наиболее употребительные слова мужской устной речи в зависимости от коммуникативной ситуации (стиля речи)

Бытовая			Профессиональная			Образовательная			«Клиент-сервис»		
Ранг	**Слово**	**%**	**Ранг**	**Слово**	**%**	**Ранг**	**Слово**	**%**	**Ранг**	**Слово**	**%**
1	ну	2,53	1	вот	2,77	1	это	4,15	1	вот	2,89
2	я	2,35	2	ну	2,74	2	вот	3,29	2	ну	2,54
3	не	2,32	3	да	2,57	3	я	3,16	3	в	2,18
4	в	2,12	4	я	2,49	4	да	2,44	4	я	2,01
5	вот	2,05	5	а	2,42	5	не	2,11	5	там	1,95
6	там	1,87	6	не	2,33	6	а	2,00	6	да	1,89
7	да	1,81	7	и	1,87	7	(э)	1,94	7	не	1,81

Бытовая			Профессиональная			Образовательная			«Клиент-сервис»		
Ранг	**Слово**	**%**	**Ранг**	**Слово**	**%**	**Ранг**	**Слово**	**%**	**Ранг**	**Слово**	**%**
8	и	1,81	8	в	1,72	8	и	1,87	8	(э)	1,75
9	а	1,77	9	у	1,65	9	что	1,66	9	что	1,72
10	что	1,61	10	это	1,46	10	ну	1,45	10	это	1,70
11	на	1,43	11	что	1,43	11	всё	1,16	11	а	1,68
12	это	1,40	12	там	1,41	12	нет	1,12	12	у	1,56
13	у	1,12	13	так	1,32	13	на	1,10	13	и	1,56
14	так	0,97	14	на	1,13	14	так	1,05	14	так	1,43
15	ты	0,91	15	как	1,05	15	там	1,01	15	всё	1,04
16	он	0,87	16	то	1,04	16	есть	0,97	16	угу	1,04
17	(э)	0,78	17	всё	0,82	17	он	0,95	17	на	0,94
18	как	0,77	18	(э)	0,69	18	в	0,93	18	нет	0,94
19	то	0,76	19	меня	0,69	19	то	0,91	19	то	0,94
20	с	0,75	20	с	0,60	20	надо	0,84	20	с	0,89
21	всё	0,75	21	сейчас	0,58	21	здесь	0,78	21	он	0,87
22	бл**ь	0,67	22	нет	0,58	22	понятно	0,78	22	как	0,73
23	нет	0,60	23	он	0,57	23	сейчас	0,76	23	есть	0,67
24	есть	0,56	24	они	0,53	24	у	0,69	24	сейчас	0,62
25	они	0,48	25	угу	0,51	25	как	0,63	25	надо	0,54
26	меня	0,46	26	же	0,47	26	ты	0,61	26	будет	0,50
27	сейчас	0,44	27	значит	0,45	27	ага	0,59	27	мне	0,46
28	мне	0,44	28	бы	0,41	28	идёт	0,57	28	хорошо	0,46
29	просто	0,42	29	говорит	0,41	29	угу	0,51	29	если	0,46
30	уже	0,42	30	уже	0,41	30	по	0,49	30	вам	0,44
31	но	0,41	31	мне	0,41	31	она	0,49	31	они	0,44
32	ещё	0,40	32	есть	0,40	32	будет	0,44	32	уже	0,42
33	блин	0,40	33	ещё	0,40	33	значит	0,42	33	вы	0,42
34	она	0,39	34	она	0,40	34	с	0,42	34	же	0,42
35	же	0,38	35	за	0,37	35	просто	0,42	35	но	0,40
36	по	0,38	36	нас	0,37	36	уже	0,40	36	меня	0,39
37	вообще	0,37	37	блин	0,37	37	теперь	0,38	37	по	0,39
38	надо	0,35	38	ага	0,35	38	для	0,36	38	ещё	0,37
39	значит	0,34	39	просто	0,35	39	знаю	0,36	39	за	0,35
40	если	0,33	40	по	0,34	40	меня	0,34	40	ты	0,35
41	бы	0,33	41	ты	0,33	41	мне	0,34	41	значит	0,33
42	знаю	0,32	42	было	0,33	42	или	0,34	42	можно	0,33
43	короче	0,29	43	вы	0,33	43	его	0,34	43	вас	0,31
44	такой	0,28	44	тоже	0,32	44	тебе	0,34	44	или	0,31
45	мы	0,27	45	говорю	0,32	45	мы	0,32	45	может	0,29
46	потом	0,27	46	тут	0,31	46	тут	0,32	46	(м)	0,27
47	говорит	0,27	47	мы	0,30	47	но	0,32	47	семь	0,27
48	все	0,27	48	надо	0,29	48	когда	0,32	48	потом	0,25
49	тебе	0,26	49	но	0,29	49	если	0,30	49	просто	0,25
50	тут	0,25	50	если	0,29	50	самое	0,30	50	тоже	0,25

Приведенные данные показывают, что неформальная устная речь характеризуется бóльшей спецификой по сравнению с тремя другими рассматриваемыми стилями. Проявляется это, в частности, в бóльшем разнообразии в верхней зоне частотного словаря личных местоимений (особенно в женской речи), дискурсивных единиц (*ну*, *короче*, *вообще* и др.), а также в использовании в мужской речи непечатной лексики. С другой стороны, в бытовой речи обеих полов реже встречаются хезитации, «речевая поддержка» *угу*, дискурсивный маркер *вот* и такие слова, как *хорошо*, *сейчас, значит*.

6. Заключение: О состоятельности полученных списков и перспективах анализа

Поскольку сопоставимых корпусов повседневной устной русской речи не существует (что в большой степени объясняется трудоемкостью как сбора, так и обработки «живого» речевого материала), валидация полученных частотных списков в настоящее время затруднена. Для оценки их состоятельности можно предложить привлечение других, близких по объему, подвыборок рассматриваемого корпуса ОРД. Такие последовательные независимые выборки позволят оценить, при каком минимальном объеме выборочного наблюдения (количество говорящих, количество слов) в районе какого ранга наступит стабилизация верхней зоны частотного словаря [Мартыненко, 1988].

Проделанная работа носит в некоторой степени иллюстративный характер, показывая возможности исследования речевых данных с помощью специальным образом аннотированного корпуса. Тем не менее, можно полагать, что и данные списки наиболее употребительных словоформ и полученные частотные статистики, особенно в верхней зоне и для общего словаря, мужского и женского словарей в целом и их бытовой речи, на данном этапе являются удовлетворительной аппроксимацией дистрибуции наиболее употребительных словоформ, характерной для повседневной устной русской речи.

Более достоверные данные будут получены позднее в результате обработки бóльших объемов корпуса ОРД. В частности, планируется построение общего словаря повседневной речи на 1 млн словоформ, а также ряда словарей для разных социальных групп говорящих (гендерных, возрастных, профессиональных).

Литература

1. *Asinovsky A., Bogdanova N., Rusakova M., Ryko A., Stepanova S., Sherstinova T.* (2009), The ORD Speech Corpus of Russian Everyday Communication “One Speaker’s Day”: Creation Principles and Annotation, Proc. 12th Int. Conf. TSD 2009, LNAI, vol. 5729, Springer, Berlin-Heidelberg, pp. 250–257.
2. *Baker P.* (2010), Sociolinguistics and Corpus Linguistics, Edinburgh University Press, Edinburgh.

3. *Bogdanova-Beglarian N., Sherstinova T., Martynenko G.* (2015), The "One Day of Speech" Corpus: Phonetic and Syntactic Studies of Everyday Spoken Russian, Proc. 18th Int. Conf "Speech and Computer" (SPECOM-2015), LNAI, vol. 9319, Springer, Switzerland, pp. 429–437.
4. *Burnard L.* (ed.) (2007), Reference guide for the British National Corpus (XML edition). Published for the British National Corpus Consortium by Oxford University Computing Services, available at: http://www.natcorp.ox.ac.uk/docs/URG/
5. *Campbell N.* (2004), Speech & Expression; the Value of a Longitudinal Corpus, LREC 2004, Lisbon, pp. 183–186.
6. *Coupland N.* (2007), Style: Language Variation and Identity, Cambridge University Press: Cambridge.
7. *Lakoff R.* (1975), Language and Woman's Place, Harper and Row, New York.
8. *Leech G., Rayson P., Wilson A.* (2001), Word Frequencies in Written and Spoken English: based on the British National Corpus, Longman, London.
9. *Popescu, I.-I.* (2009), Quantitative Linguistics: Word Frequency Studies, Mouton de Gruyter, Berlin-New-York.
10. *Rayson P., Leech G., Hodges M.* (1997), Social differentiation in the use of english vocabulary: some analyses of the conversational component of the British National Corpus, International Journal of Corpus Linguistics, 2 (1), pp. 133–152.
11. *Romaine S.* (2008), Corpus linguistics and sociolinguistics, Corpus Linguistics: An International Handbook, Mouton de Gruyter, Berlin-New York, vol. 1, pp. 96–111.
12. *Sherstinova, T.* (2015) Macro Episodes of Russian Everyday Oral Communication: towards Pragmatic Annotation of the ORD Speech Corpus / Ronzhin, A. et al. (eds.) SPECOM 2015, Lecture Notes in Artificial Intelligence, LNAI, vol. 9319. Springer, Switzerland, pp. 268–276
13. *Stenström, A.-B.* (1991), Expletives in the London-Lund Corpus, English Corpus Linguistics in Honour of Jan Svartvik, Longman, London, pp. 230–253.
14. *Tannen D.* (1991), You Just Don't Understand: Women and Men in Conversation. Virago Press, London.
15. *Ляшевская О. Н., Шаров С. А., Частотный словарь современного русского языка* (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009. http://dict.ruslang.ru/freq.php
16. *Мартыненко Г. Я.* (1988) Основы стилеметрии. Л.: ЛГУ.
17. *Потапова Р. К., Потапов В. В.* (2006) Язык, речь, личность. М.: Языки славянской культуры. 496 с.
18. *Шайкевич А. Я.* (2015) Меры лексического сходства частотных словарей, Корпусная лингвистика — 2015. Труды международной конференции. Ответственные редакторы: Захаров В. П., Митрофанова О. А., Хохлова М. В. С. 422–429.

References

1. *Asinovsky A., Bogdanova N., Rusakova M., Ryko A., Stepanova S., Sherstinova T.* (2009), The ORD Speech Corpus of Russian Everyday Communication "One

Speaker's Day": Creation Principles and Annotation, Proc. 12th Int. Conf. TSD 2009, LNAI, vol. 5729, Springer, Berlin-Heidelberg, pp. 250–257.
2. *Baker P.* (2010), Sociolinguistics and Corpus Linguistics, Edinburgh University Press, Edinburgh.
3. *Bogdanova-Beglarian N., Sherstinova T., Martynenko G.* (2015), The "One Day of Speech" Corpus: Phonetic and Syntactic Studies of Everyday Spoken Russian, Proc. 18th Int. Conf "Speech and Computer" (SPECOM-2015), LNAI, vol. 9319, Springer, Switzerland, pp. 429–437.
4. *Burnard L.* (ed.) (2007), Reference guide for the British National Corpus (XML edition). Published for the British National Corpus Consortium by Oxford University Computing Services, available at: http://www.natcorp.ox.ac.uk/docs/URG/
5. *Campbell N.* (2004), Speech & Expression; the Value of a Longitudinal Corpus, LREC 2004, Lisbon, pp. 183–186.
6. *Coupland N.* (2007), Style: Language Variation and Identity, Cambridge University Press: Cambridge.
7. *Lakoff R.* (1975), Language and Woman's Place, Harper and Row, New York.
8. *Leech G., Rayson P., Wilson A.* (2001), Word Frequencies in Written and Spoken English: based on the British National Corpus, Longman, London.
9. *Lyashevskaya O. N., Sharov S. A.* (2009), Frequency List of Modern Russian language (on the Materials of the Russian National Corpus) [Chastotnyj slovar' sovremennogo russkogo yazyka (na materialah Nacional'nogo korpusa russkogo yazyka)], Azbukovnik, Moscow, available at: http://dict.ruslang.ru/freq.php
10. *Martynenko G. Ya.* (1988), Foundations of Stylometrics [Osnovy stilemetrii], Leningrad State University, Leningrad.
11. *Popescu, I.-I.* (2009), Quantitative Linguistics: Word Frequency Studies, Mouton de Gruyter, Berlin-New-York.
12. *Potapova R. K., Potapov V. V.* (2006), Language, speech, personality [Yazyk, rech', lichnost'], Yazyki slavyanskoj kul'tury, Moscow.
13. *Rayson P., Leech G., Hodges M.* (1997), Social differentiation in the use of english vocabulary: some analyses of the conversational component of the British National Corpus, International Journal of Corpus Linguistics, 2 (1), pp. 133–152.
14. *Romaine S.* (2008), Corpus linguistics and sociolinguistics, Corpus Linguistics: An International Handbook, Mouton de Gruyter, Berlin-New York, vol. 1, pp. 96–111.
15. *Shajkevich A. Ya.* (2015), Measures of lexical similarity between frequency dictionaries [Mery leksicheskogo skhodstva chastotnyh slovarey], Proc. of the Int. Conference "Corpus linguistics-2015" [Trudy mezhd. konf. "Korpusnaya linguistica-2015"], St. Petersburg State University, St. Petersburg, pp. 422–429.
16. *Sherstinova, T.* (2015), Macro episodes of Russian everyday oral communication: towards pragmatic annotation of the ORD speech corpus, Proc. 18th Int. Conf. "Speech and Computer" (SPECOM-2015), LNAI, vol. 9319, Springer, Switzerland, pp. 268–276.
17. *Stenström, A.-B.* (1991), Expletives in the London-Lund Corpus, English Corpus Linguistics in Honour of Jan Svartvik, Longman, London, pp. 230–253.
18. *Tannen D.* (1991), You Just Don't Understand: Women and Men in Conversation. Virago Press, London.

MULTI-PRONUNCIATION LEXICON FOR RUSSIAN AUTOMATIC SPEECH RECOGNITION (PILOT STUDY)

Shirokova A. (anna_a@stel.ru),
Telesnin B. (telesnin_ba@stel.ru),
Rogozhina V. (mind_your_own_business@rambler.ru)

Stel CS, MSLU, Moscow, Russia

Our pilot study is aimed at building a lexicon of effective pronunciation variants on the basis of canonical pronunciations, for implementing it into the automatic speech recognition system for Russian. We focus on phonetic changes in word pronunciation caused by different factors operating in spontaneous speech. Our speech data includes three different corpora of the conversational type. Manual expert processing and analysis of the audio data are used. The lexicon construction procedure is given. Some statistics for pronunciation variation in Russian, obtained from the speech data, is presented. A description of frequent types of this phenomenon is given. Parallel and sequential pronunciation variants are discussed. Ways of formulating general phonetic variation rules and predicting potential contexts, in which pronunciation variation is likely to appear, are considered. Test data, phoneset used, and automatic speech recognition (ASR) parameters are described. Preliminary results for ASR and key word spotting (KWS) are shown. The appropriateness of using multi-pronunciation lexicon is discussed.

Keywords: Russian spontaneous speech, pronunciation variants, Russian pronunciation, spontaneous speech, pronunciation lexicon, reduction, Russian ASR

1. Introduction

Variations in word pronunciation have multiple sources. First, it is a common phenomenon across languages that words share the same written form but have different pronunciations (homographs). Also there are orthoepic ambiguities when a word has multiple pronunciations which are orthoepically acceptable. Specific pronunciations reveal and are due to individual manner or regional accent of a speaker. It is generally known that speech genres and speaking styles determine pronunciation peculiarities. In particular, relaxed or condensed pronunciation typical for rapid fluent (especially informal) spontaneous speech is characterized by various forms of contractions, reductions, elisions, deletions, etc. The above processes drastically affect articulatory and acoustic parameters of phones and cause grave changes in sound image of a word.

Pronunciation disambiguation is essentially important in speech synthesis, speech recognition and other fields of automatic natural language processing. Most state-of-the-art

ASR systems use phone-based representations for acoustic modeling. As stated in (Schultz, Kirchhoff 2006), explicitly specified pronunciations allow spoken language to be modeled more accurately. A pronunciation-based approach includes the potential for reducing the ambiguity of a given language writing system. If different acoustic realizations of a word are unlikely to be covered properly by the acoustic models, a given lexical entry may be assigned multiple pronunciations to represent these significant differences. When adding variants, one has to consider the types of speech that will be processed in order to add pronunciation variants relevant for the actual genre and style.

Our work is aimed at constructing a lexicon of effective pronunciation variants on the basis of the canonical pronunciations and implementing it into the ASR system for Russian (Zulkarneev&al 2013). We take preliminary ASR and KWS experiments to roughly assess a potential profit of the explicit adding of phonetic variants for reduced tokens. Furthermore, our study is intended to assess the very appropriateness of taking into account multiple pronunciations in our ASR projects. It is essential to analyze whether there exist trends towards ASR performance gain achieved by using such an enhanced lexicon.

2. Related work

Our project has been inspired by a series of researches that deals with pronunciation variation phenomena and its influence on automatic speech recognition. The work (Adda-Decker, Lamel 1998) is aimed at evaluating the use of pronunciation variants across different system configurations, languages (English and French) and speaking styles (spontaneous and read speech). A correlation between the word frequency and the number of productive variants is outlined.

In (Adda-Decker&al 1999) authors focus on well-known pronunciation variants in French: the so-called mute *e* and liaisons. Their frequencies of occurrence in read speech and spontaneous speech are computed and compared regarding these types of speech.

Formal phonetic rules are recently developed for Austrian German conversational speech (Schuppler&al 2014).

For the Russian language the issue of pronunciation variety has been studied in theoretical and applied aspects.

The monograph (Bondarko&al 1988) describes the phonetic system of spontaneous speech. It is based mainly on evidences of the Russian oral speech, but also considers some English and German data. It covers the problems of the pronunciation norm and acceptable variation, the allowable phonetic realization of phonemic units and the pronunciation types. As a supplement it includes a phonetic lexicon of 80 Russian high frequency words which gives a number of different phonetic representations for each word.

In (Lobanov, Tsirul'nik 2007) it is claimed that it is possible to predict potential contexts in which pronunciation variation is likely to appear in conversational speech. Moreover, there are deduced systematic phonetic changes in word pronunciation caused by the above factors which are generalized and formulated as strict phonetic rules.

An algorithm for automatic generation of pronunciation variants for Russian based on the results of the above-mentioned research (Lobanov, Tsirul'nik 2007)

is proposed in (Kipyatkova, Karpov 2009) and is reported to be implemented into Russian ASR system (Kipyatkova&al 2013).

The pronunciation variety and peculiarities of reduced word forms in the ORD speech corpus of Russian everyday communication are analysed in (Bogdanova, Palshina 2010).

3. Pronunciation variants

3.1. Speech Data

Our speech data involve three separate speech corpora. The first corpus of retrieval queries contains short utterances of more than 4000 speakers of different gender and age. Speech material includes mostly exact address requests, geographic objects requests and proper names requests. Another corpus of professional telephone speech contains recordings of power engineers professional conversations. There are 30 adult male speakers. It is characterized by high portion of professional lexis, proper names and toponyms. The third corpus contains telephone speech recordings of the general conversational type. In all the corpora there is a portion of speakers with more or less distinct regional accent features. Table 1 summarizes the corpora characteristics. The given conversational data represents rapid fluent spontaneous speech. All the corpora are not publically available and are the property of our customers.

Table 1. Corpora characteristics

Corpus	Queries	Professional	Conversational
Number of speakers (multiple records per speaker available)	4,000+	30	1,000+
Duration average (per record)	10 sec	3 min	5 min
Duration total	30 h+	10 h+	50 h+
Gender	all	male	all
Age	all	adult	all

3.2. Building Canonical Pronunciations

Although there are many grapheme-to-phoneme conversion techniques (Bisani, Ney 2008), in our work we use a rule-based automatic transcription system to build the canonical pronunciations of words in Russian. This multifunctional transcribing tool (Krivnova&al 2001) has different representation levels: phonemes, phones, etc. Context-dependent rules cover major and slight conversion patterns. The system uses a number of exception lists. The phoneme error rate is 2% (mostly in proper names and loanwords) when testing on 500 Kb of texts.

3.3. Building Pronunciation Variants

The annotations of the speech data were created by linguists according to the adopted guidelines (Glavatskih&al 2015). A multi-pronunciation lexicon has been created on the basis of the orthographic transcripts which includes both the canonical phonetic representations of words and their pronunciation variants. The latter were created manually by expert phoneticians relying on the results of the perceptive and acoustic analyses. The process of building a lexicon had the following steps:

- when creating orthographic transcripts of the speech data expert phoneticians were asked to mark words with gravely reduced pronunciations (contracted forms, phone deletion) and incorrect or non-standard pronunciations (stress position, etc);
- marked words were ranked according to their frequency of occurrence in each data set; lists of the most frequent words were taken into account;
- up to four most common variants of actual pronunciations were added to each lexicon (after speech fragments corresponding to the marked words had been listened to by experts).

Thus, three separate lexicons have been formed. It should be noted that since those were not simultaneous projects, the data of the previously built lexicon was taken into account when creating a new one. It is obvious that they must share some lexical entries, but due to the corpora specifics their set of variants can still be partly different. Conversational lexicon has been chosen for the further processing, since the others contain a high portion of specific lexical data such as proper names, toponyms, etc.

3.4. Pronunciation Lexicon for Conversational Corpus

The following Table 2 and Table 3 highlight the main tendency in word reduction and its usage within 50 hours of spontaneous speech. The ratio of total amount of reduced words to total amount of words used in the database equals 5.53 % (see Table 2). It should be noted, however, that only word tokens with evident reduction (phone deletion, syllable contraction) are treated as reduced variants, other segmental changes are supposed to be covered by the acoustic models. Although, as an observation, such small reduction ratio in total corpora implicitly testifies, as we see it, that the impact of adding reduced tokens for the total lexicon is not significant and, therefore, can be disregarded.

Table 2. Conversational corpus statistics

Duration of spontaneous speech	50 hours
Total amount of pronounced words	228,209
Total amount of reduced words	12,611
Ratio of total amount of reduced words to total amount of pronounced words, %	5.53

The list of the most frequent words affected by reduction is given in Table 3.

Table 3. List of the most frequent reduced words in conversational corpus

Reduced words	Frequency of reduced realizations	Overall frequency of the word	Ratio of reduced realizations to overall frequency of the word, %
что (what)	3,386	5,780	58.58
сейчас (now)	1,822	2,005	90.87
тогда (then)	561	987	56.84
сегодня (today)	511	779	65.60
говорить (to say) (all the verb forms are taken into account)	943	2,112	44.65
ничего (nothing)	427	632	67.56
чтоб (in order to)	480	777	61.78
только (just)	365	550	66.36
тебе (for you)	306	1,052	29.09
алло (hello)	322	845	38.11
сколько (how many)	232	357	64.99
когда (when)	231	491	47.05
тебя (you)	203	646	31.42
наверное (perhaps)	212	264	80.30
здравствуйте (greetings)	103	230	44.78

3.5. Phonetic evidence

As it has been suggested in (Adda-Decker, Lamel 1998) we use parallel (equipollent) and sequential (derived) pronunciation variants. Parallel are predominantly used to cover homographs and orthoepically acceptable variants, while sequential represent different stages of reduction.

The examples given in Table 5 below and rules formulated in (Lobanov, Tsirul'nik 2007), (Kipyatkova, Karpov 2009) show that reduction can be viewed as a categorical phenomenon which can lead to the change of one phonological feature into another or to the total deletion of different segments.

According to (Hoole&al 2012) a lot of Russian words in spontaneous speech tend to syncope, (i.e. the strategy to make trisyllabic word bisyllabic), for example as in word 'сегодня' ('today') shown in Figure 1 in its full pronunciation. Figure 2 displays almost the deletion of the first syllable vowel [i], whereas there are some [i]-traces in the formant curve. There is no surprise in deletion of the phone [v] due to its intervocal position, while the optional presence of the phone [d^j] is due to the potential total regressive assimilation to [n^j], which has the same place of articulation.

Tokens can be contracted even more, and this illustrates the gradient character of reduction process operating to some sequence of segments.

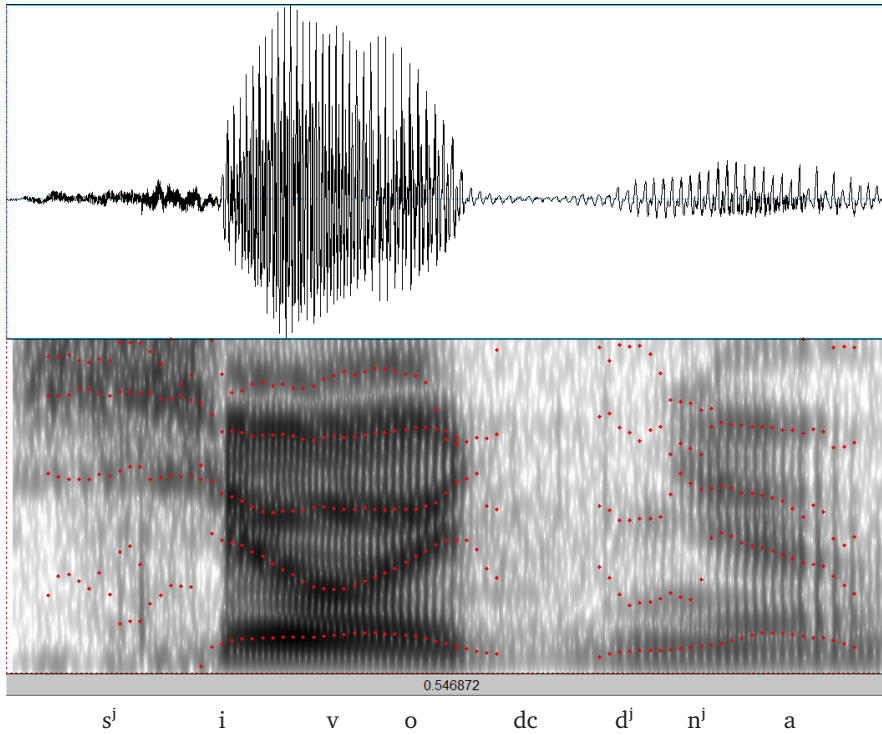


Fig. 1. Oscillogram and spectrogram for a pronunciation of the word ‘сегодня’ [sʲ i v o dʲ nʲ a]. Figures are captured within our software annotation tool described in (Glavatskih&al 2015)

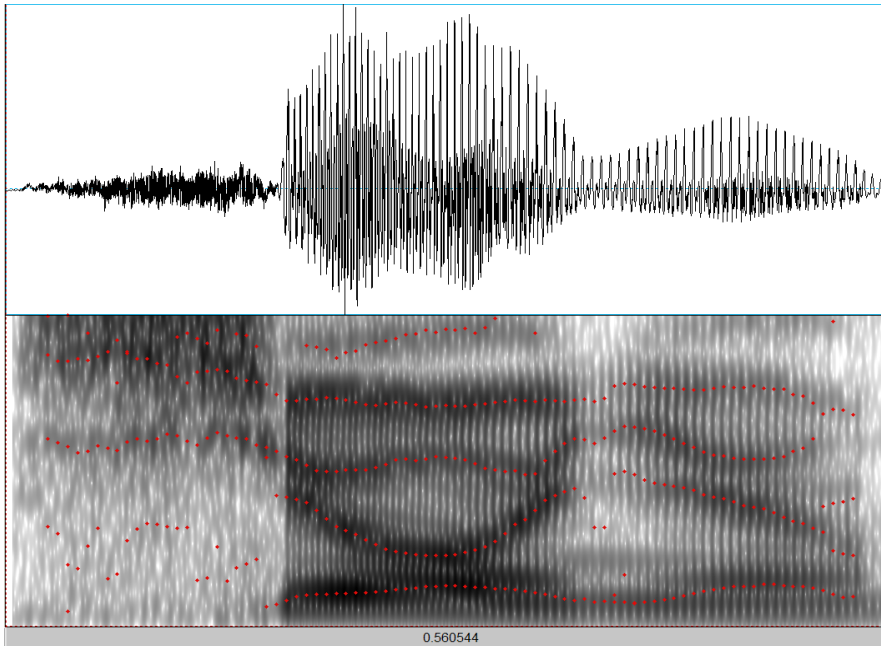


Fig. 2. Oscillogram and spectrogram for a pronunciation of the word 'сегодня' [sʲ o nʲː ə]

3.6. General variation types

Our general observations verify that vowels are more robust to duration and quality reduction as well as to deletion than consonants. Experts have marked phonetic changes that are reported in (Bondarko&al 1988) to be systematic and typical for spontaneous speech. Those referred to consonants include:

- deletion of /j/ in word initial, word final, intervocal positions and in V/j/C contexts;
- deletion of /v, vʲ, bʲ, dʲ/ in intervocal position;
- deletion of one of double consonants;
- deletion of word-final plosives;
- consonant cluster reduction including phonetic changes across word boundaries (strong assimilation or total deletion of phones and phone sequences);

For vowels the following observations are made:

- stronger duration reduction even in stressed syllables;
- quality reduction in unlike position to the stressed syllable;
- delabialization of /u, o/ in weak positions;
- quality reduction of /u, y/ in weak positions;
- centralization in weak position;
- vowel deletion in unstressed syllables.

It is verified by all our data that numerals, common words and notional word fillers are the first to suffer compression in fluent spontaneous speech.

4. ASR & KWS Experiments

4.1. Baseline & Training Data

The experiments were performed using a speech recognition system based on Kaldi (Povey&al 2011). Recognition was carried out in two stages, and it can be described as hybrid HMM-DNN approach.

The training data (68.2 hours) is based on multiple sources: the major part of the data listed above, our Broadcast Russian speech data (Glavatskih&al 2015) and Russian Voxforge open speech corpus.

At first stage recognition system based on HMM was used to build the adaptation of MLLR matrix, then the MLLR transformation was applied to feature vectors. At the second stage recognition was performed using DNN, consisting of 4 hidden layers, each hidden layer composed of 1,024 elements.

As the language model, 3-gram model was used in speech recognition, trained on the text data of 772,365 words within 114,423 phrases, its pronunciation lexicon equals 44,446 tokens. Phrases, that are heavily distorted by artifacts due to the channel and/or other technical issues, or do not contain any intelligible word, are filtered and not used for the further analysis. Thus only 98,331 utterances (640,242 pronounced words and 38,737 word tokens), which equals approximately 54 hours of spontaneous speech, are used for building the acoustic models. Acoustic model adaptation set is composed of 38 hours (34,861 utterances, 322,861 pronounced words and 24,582 word tokens) of spontaneous speech. Speaker-independent model is applied.

4.2. Test Data

In our system the phonetic alphabet Worldbet is used, though our set of symbols for the Russian language slightly differs from the one suggested in (Hieronymus 1993).Thus, the property “dental” is not marked in plosives, nasals and the affricate. Furthermore, the symbol “I” is added to the vowel inventory. It represents the reduced high front vowel (the second stage of reduction) which is not considered in (Hieronymus 1993). The symbol “ax” represents the reduced mid central vowel (the second stage of reduction) and corresponds to “&” in the Worldbet list. It should be noted, that in Russian phonetic transcription palatalization is only marked when there is a corresponding nonpalatalized consonant. In unpaired consonants this property is not marked, since it is supposed to be implied in the symbol itself. “Ix” represents the reduced diphthongoid that has a higher and narrower beginning and a wider and lower ending. It is regarded as a prototypic realization of post-stressed

combinations of /j/ or /i/ and a wider and lower vowel. The vowel reduction in the terminal open post-stressed syllables is supposed to be only slight and for that reason is not taken into account. In continuous speech, however, terminal post-stressed vowels are reduced according to general rules. The above phoneset is mapped to IPA in Table 4.

Table 4. Phoneset mapping to IPA

IPA	Our Phoneset
ɪə	Ix
ʂ	S
ɕ	Sj
ʐ	Z
a	a
ə	ax
b	b
b^j	bj
d	d
voiced closure	dc
d^j	dj
e	e
f	f
f^j	fj
g	g
g^j	gj
i	i
ɨ	ix
j	j
k	k
k^j	kj
ɬ	l
l^j	lj

IPA	Our Phoneset
m	m
m^j	mj
n	n
n^j	nj
o	o
p	p
p^j	pj
r	r
r^j	rj
s	s
s^j	sj
t	t
t͡s	ts
t͡ɕ	tSj
unvoiced closure	tc
t^j	tj
u	u
v	v
v^j	vj
x	x
x^j	xj
z	z
z^j	zj

We take preliminary speech recognition and key word spotting experiments to analyze the potential of the performance improvement when taking into account pronunciation variation of reduced words. Speech data for the test (not included in the training set) is 2 hours of spontaneous speech, i.e. 1,591 utterances with 15,836 words and 3,324 word tokens. In the course of experiment 33 most frequently used words were selected and their pronunciation variants were processed. Table 5 covers the majority of pronunciation variants built for several words included in the test set. The upper transcriptions for each word given in Table 5 represent these words as pronounced solely.

Table 5. Pronunciation variants for frequently used words

Frequent words	Pronunciation variants
сейчас (now)	sj i tc tSj a s Sj a s Sj a
что (what)	S tc t o S tc t ax S o tc tSj o
говорит (says)	dc g ax v a rj i tc t dc g ax a rj i tc t dc g a rj i tc t dc g rj i tc t dc g I tc t
тебе (for you)	tc tj i dc bj e tc tj i e tc tj e
когда (when)	tc k a dc g dc d a tc k a dc d a
сегодня (today)	sj i v o dc dj nj a sj i o dc dj nj ax sj o dc dj nj a sj o nj ax
будет (will)	dc b u dc dj I tc t dc b u I tc t dc b u I dc b u tc t
сказала (said)	s tc k a z a l a s tc k a a l ax s tc k a l ax
двадцать (twenty)	dc d v a tc tc ts ax tctj dc d v a tc tKs&
позвонишь (you'll call)	tc p ax z v a nj i S tc p a z v o nj I S

4.3. Preliminary Results & Discussion

Three tests based on the same acoustic and language models are carried out. The tests differ in pronunciation variants that have been used. For test_result_1tr only one canonical pronunciation variant is applied, so it represents baseline. For test_result_2tr only one additional variant is included, whereas in test_result_vartr four pronunciation variants are added (when given, otherwise less). Preliminary results are obtained (see Table 6), where FA and FR denote the false acceptance and the false rejection rates.

As a result of adding multi-variant pronunciation, the system manages to detect the word tokens that differ from their standard pronunciation. For that reason the

correctness rate (CORR) increases. On the other hand, when all the pronunciation variants are taken into account, as test_result_vartr shows, it leads to the increase of word error rate (WER) which is due to a higher number of insertions.

Table 6. System performance for ASR and KWS

tests	CORR%	WER%	FA	FR%
Baseline	64.01	40.44	2.005	26.98
Test_result_2tr	64.16	40.31	2.33	25.57
Test_result_vartr	64.31	40.62	3.52	22.67

It is supposed that strongly reduced variants tend to appear when an acoustic observation is unlikely to be recognized adequately. To reduce the number of insertions and, as a consequence, WER, the shortest transcriptions should be eliminated from the pronunciation lexicon and another set of experiments needs to be taken. For such purpose a special technique should be applied, which would enable to track the actual system choice of a pronunciation variant in the recognition process.

5. Conclusion

Our research verifies, expands and specifies the experimental results shown at (Bondarko&al 1988). It has been observed that the most likely words to be affected by reduction and adjacent mechanisms are numerals, common words and notional word fillers, which seems reasonable in the context that frequent words carrying little information are drastically affected by articulation relaxation. Obviously, the most robust segments of the words are stressed vowels, while in weak especially post-tonic syllables phone and phone sequences deletion is likely to appear. At the current stage of work there can be outlined the major factors that account for the evidences of pronunciation variation in the speech corpora, however, the potential contexts and the actual conditions can hardly be described in terms of patterns and summarized as a set of phonetic rules.

As it has been noted the pronunciation variants in our lexicon were created manually. Moreover, experts had to listen to audio samples in order to specify exact pronunciation and then to select the most common ones. Unfortunately, manual processing did not allow us to assign a unique acoustic form to its specific phonetic representation. This partly explains the lack of statistical data and estimations for the phone deletion and other segmental changes.

It is worth mentioning that experts are limited by the phone set of the speech recognition system so they have to approximate their actual observation and refer it to some phonetic unit in the set, therefore missing some slight phonetic differences.

Our further research involves learning potential contexts of the phonetic changes localization to be able to predict their occurrence. Alongside with it we plan to investigate the character of phonetic changes and to systematize their types regarding contexts of appearance.

As stated in (Adda-Decker, Lamel 1998), one should be aware that different words are likely to share the same pronunciation variants, when applying such multi-lexicon to the ASR system. It inevitably leads to a higher rate of word confusion.

In the further research the following strategies can be implemented to effectively introduce multi-pronunciation lexicon to ASR:

- to train acoustic models within added pronunciation variants to provide model compatibility;
- to take into account less variants and to select the most effective ones;
- to assign weight to variants.

Still, the question is open about the optimal strategy of such lexicon building, whether it is worth deducing general phonetic modification rules and applying them totally to all the words in lexicon (to a part of it) as suggested in (Kipyatkova, Karpov 2009), or to manually provide selected words with required pronunciations. The latter is not always the slowest strategy (if a list is not large), but enables including specific actual pronunciations and controlling their confusion potential. For the same reason the number of pronunciation variants should be limited. In any case the appropriateness of implementing an enhanced lexicon should be studied more thoroughly and the supposed performance profit should be estimated more accurately.

References

1. *Adda-Decker M., Lamel L.* (1998), Pronunciation variants across systems, languages and speaking style in Proc. ESCA Conf., May 1998, pp. 1–6.
2. *Adda-Decker M., Boula de Mareüil P., Lamel L.* (1999), "Pronunciation variants in French: schwa & liaison", in Proc. ICPhS Conf., 1999, pp. 2239–2242.
3. *Bisani M., Ney H.* (2008), Joint-sequence models for grapheme-to-phoneme conversion, Speech Communication, vol.50, May 2008, pp. 434–451.
4. *Bogdanova N. V., Palshina D. A.* (2010), The reduced forms in Russian (a lexicographical description) in Proc. Word. Lexicon. Philology: Lexicon Text and Lexicographical Content Conf., Nov. 2010, pp. 491–497.
5. *Bondarko L. V., Verbitskaja L. A., and N. I. Geilman* (1988) Spontaneous Speech Phonetics. [Fonetika spontannoy rechi] St. Petersburg: 1988.
6. *Glavatskih I. A., Platonova T. S., Rogozhina V. S., Shirokova A. M., Smolina A. A., Kotov M. A., Ovsyannikova A. S., Repalov S. A., Zulkarneev M. Ju.* (2015), The multi-level approach to speech corpora annotation for automatic speech recognition, in Proc. SPECOM Conf., Sept. 2015, pp. 438–445.
7. *Hieronymus J. L.* (1993) "ASCII phonetic symbols for the world's languages: Worldbet, Journal of International Phonetic association, Vol. 23.
8. *Hoole P., Bombien L., Pouplier M., Mooshammer C., Kuhnert B.* (2012), Consonant Clusters and Structural Complexity. Munich: Mouton de Gruyter.
9. *Kipyatkova I., Karpov A.* (2009), Creation of multiple word transcriptions for conversational Russian speech recognition, in Proc. SPECOM Conf., Sept. 2009, pp. 71–75.

10. *Kipyatkova I., Karpov A., Verkhodanova V., Zelezny M.* (2013), Modeling of pronunciation, language and nonverbal units at conversational Russian speech recognition, International Journal of Computer Science and Applications, vol. 10, №1, 2013, pp. 11–30.
11. *Krivnova O. F., Zakharov L. M., Strokin G. S.* (2001), Automatic transcriber of Russian texts: problems, structure and application in Proc. SPECOM Conf., Sept. 2001, pp. 408–409.
12. *Lobanov B. M., Tsirul'nik L.I.* (2007), Modelling of in-word and word-boundary phonetic-acoustic phenomena in full and conversational speaking styles for TTS synthesizer MULTIFON [Modelirovanie vnutrislovnykh i mezhslovnykh fonetiko-akusticheskikh yavleniy polnogo i razgovornogo stiley rechi v sisteme sinteza rechi po tekstu MULTIFON] in Proc. The first interdisciplinary seminar "Russian conversational speech analysis" RCSA [Trudy pervogo mezhdistsiplinarnogo seminara "Analiz russkoy razgovornoy rechi"], St-Peterburg, — Spb.: GUAP, Aug. 2007, c. 57–71.
13. *Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlicek P., Qian Y., Schwarz P., Silovsky J., Stemmer G., Vesely K.* (2011), The Kaldi speech recognition toolkit in Proc. IEEE Conf., 2011, iEEE Catalog No.: CFP11SRW-USB.
14. *Schultz T., Kirchhoff K.* (2006), Multilingual Speech Processing. Elsevier.
15. *Zulkarneev M., Grigoryan R., Shamraev N.* (2013), Acoustic modeling with deep belief network for Russian speech recognition, in Proc. SPECOM Conf., Sept. 2013, pp. 17–24.
16. *Schuppler B., Adda-Decker M., Morales-Cordovilla J. A.* (2014), Pronunciation variants in read and conversational Austrian German in Proc. INTERSPEECH Conf., Sept. 2014, pp. 1453–1457.

БЕЛАРУСЬ VS. *БЕЛОРУССИЯ*: СТРУКТУРА ОДНОГО ЛИНГВОПОЛИТИЧЕСКОГО КОНФЛИКТА В СОЦИАЛЬНЫХ МЕДИА

Сомин А. А. (somin@tut.by)

Национальный исследовательский университет «Высшая школа экономики», Москва, Россия

Полий А. А. (a.a.polyi@gmail.com)

Российский государственный гуманитарный университет, Москва, Россия

В работе рассматриваются различные аспекты лингвополитического конфликта, связанного с выбором между двумя русскоязычными топонимами — *Белоруссия* и *Беларусь*, а также прилагательными *белорусский* и *беларус(с)кий* и этнонимами *белорус* и *беларус.* Суть проблемы в том, что в русском языке России используется вариант *Белоруссия*, который оценивается многими белорусами как пренебрежительный, тогда как сами белорусы используют вариант *Беларусь* в том числе и в русской речи. Для понимания структуры конфликта мы описываем, анализируя газеты 1990-х гг., как и почему появился и распространился топоним *Беларусь*, а также изучаем данные, полученные с помощью двух онлайн-опросов, количественное распределение дериватов, образованных от вариантов топонимов, и, наконец, описываем сценарии конфликтной коммуникации в комментариях в различных социальных медиа с применением качественных методов анализа. Один из опросов был направлен на определение социальной дистрибуции двух вариантов топонимов, тогда как целью другого было изучение рефлексии белорусов по отношению к топониму *Белоруссия* и его производным. В работе показано, что обе стороны конфликта имеют ограниченные наборы аргументов, которые постоянно появляются в конфликтной коммуникации в комментариях под различными статьями в интернете.

Ключевые слова: социолингвистика, конфликт, топонимика, Белоруссия, социальные медиа, языковая рефлексия

BELARUS VS. *BELORUSSIA*: THE STRUCTURE OF A LINGUO-POLITICAL CONFLICT IN SOCIAL MEDIA

Somin A. A. (somin@tut.by)

National Research University Higher School of Economics, Moscow, Russia

Poliy A. A. (a.a.polyi@gmail.com)
Russian State University for the Humanities, Moscow, Russia

This paper studies different aspects of a linguo-political conflict concerned with choosing between two Russian toponymic variants — *Belorussia* and *Belarus'* as well as adjectives *belorusskij (Belorussian)* and *belarus(s)kij (Belarusian)* and ethnonyms *belorus* and *belarus*. The core of the problem is that in the Russian language of Russia the variant *Belorussia* is used, which is considered to be insulting by many Belarusians, who prefer to use the variant *Belarus* while speaking Russian. In an attempt to understand the structure of this conflict, we analyze how and why the toponym *Belarus* appeared and spread through the newspapers of 1990-s, study the data from two online polls and the distribution of some words derived from the two toponymic variants, and finally discuss the scenarios of conflict communication in discussions in various social media. One of the polls shows the social distribution of the two toponymic variants and the other examines the attitude of the Belarusians towards the toponym *Belorussia* and its derivates. We show that each side of the conflict has its own limited set of ideas that reappear in conflict communication in comments under different articles on the Internet.

Key words: sociolinguistics, conflict, toponymy, Belarus, social media, attitudes to language

1. Введение: изменения топонимов и их конфликтность

После распада Советского Союза на территории постсоветского пространства происходили две разновидности переименования топонимов: в первом случае городам возвращались исторические досоветские названия (*Куйбышев — Самара, Фрунзе — Бишкек*), а во втором, как пишет М. А. Кронгауз, «на смену русскому названию было предложено, по существу, это же название, но на другом языке, главном языке новой независимой страны» [Кронгауз 2013:11]: *Алма-Ата — Алматы, Киргизия — Кыргызстан, Белоруссия — Беларусь, Молдавия — Молдова* и т. п. Особыми случаями являются *Таллинн* (где речь идёт только об орфографии) и вариативность предлогов в сочетании с названием Украины: *в/на Украину, с/из Украины*.

Любая смена топонима — потенциально конфликтный акт. В случае же постсоветского пространства конфликтность усугубляется двумя аспектами: во-первых, общим языком, в который вносятся изменения, — русским — у государств-инициаторов переименования и «языковой метрополии», а во-вторых, ассоциированием предыдущих вариантов топонимов с советским строем и российским доминированием.

Особенно остро конфликтность стала проявляться с развитием социальных медиа. Если раньше конфликт, связанный с употреблением одного из двух вариантов топонима, мог развиваться только при личной коммуникации или лишь с ограниченной интерактивностью — например, в формате возмущённых писем в газеты, то сейчас триггером конфликтной коммуникации может

выступить неверный с точки зрения комментатора топоним, как употреблённый в тексте электронных СМИ, так и использованный другим пользователем в дискуссии в публичном пространстве (в социальных сетях, на форуме, в блоге) — что может немедленно вызвать эмоционально-окрашенный спор в комментариях. Подобные споры являются настолько регулярным явлением, что получили особые жаргонные наименования: *хохлосра́ч* — любой яростный спор в комментариях, связанный с Украиной, *бульбосра́ч* (от бел. *бульба* ‘картошка’) — спор, связанный с Белоруссией. При этом если в первом случае из-за актуальности украинской повестки дня выбор предлога является лишь одной из тем для споров, во втором случае абсолютное большинство споров касаются именно варианта топонима.

Переход конфликтной коммуникации в публичное онлайн-пространство дал лингвистам возможность изучать архитектуру конфликтов, ранее практически недоступную для стороннего взгляда. Комбинируя качественные и количественные методы анализа социальных медиа, имея в своём распоряжении обширный корпус конфликтных диалогов, мы можем успешно описывать различные аспекты развёртывания конфликта. В настоящей статье, применяя вышеуказанные методы в сочетании с методикой опросов, экспериментов и работы с архивными текстами, мы проанализируем конфликт, связанный с вариативностью топонима *Белоруссия / Беларусь* и родственных ему слов.

2. Появление варианта *Беларусь* и развитие вариативности

В русском языке вариант *Беларусь* — собственно, белорусскоязычное название Белоруссии, существующее с конца XIX века, — появился в 1991 году с принятием 19 сентября того же года закона N 1085-XII «О названии Белорусской Советской Социалистической Республики и внесении изменений в Декларацию Верховного Совета Белорусской Советской Социалистической Республики о государственном суверенитете Белорусской Советской Социалистической Республики и Конституцию (Основной Закон) Белорусской ССР»[1]. Наиболее важен в нём, собственно, первый пункт:

> *«1. Белорусскую Советскую Социалистическую Республику впредь называть „Республика Беларусь“, а в сокращенных и составных названиях — «Беларусь».*
> *Установить, что эти названия транслитерируются на другие языки в соответствии с белорусским звучанием».*

При всей лингвистической неоднозначности этого пункта (верно ли, что слово *Рэспубліка* также должно транслитерироваться, а не переводиться?)

[1] http://zakonby.net/zakon/62709-zakon-respubliki-belarus-ot-19091991-n-1085-xii-quotonazvanii-belorusskoy-sovetskoy-socialisticheskoy-respubliki-i-vnesenii-izmeneniy-v-deklaraciyu-verhovnogo-soveta-belorusskoy-sovetskoy.html

из него следовало, что и в русском языке должно употребляться название *Беларусь*. Это указание было направлено не только на русский язык, но и на другие языки, так как в большинстве европейских языков название страны либо транслитерировалось из русского *Белоруссия* (англ. *Byelorussia* — отсюда домен .by и *Belorussia*, фр. *la Biélorussie*, ит. *Bielorussia* и т. д.), либо переводилось по частям как '*Белая Россия*' (нем. *Weißrussland*, швед. *Vitryssland*, фин. *Valko-Venäjä*, венг. *Fehéroroszország* и т. д.). Достаточно оперативно отреагировал только английский язык: в течение последующих десятилетий топоним *Belarus* постепенно вытеснял устаревшие варианты, позже это также коснулось и названия языка и национальности. Так, Дж. Смит в предисловии к своей книге «Red Nations» пишет: «The most important of these [spelling changes or new names] is the country now almost universally known as Belarus, at the insistence of the independent state's political leaders in the early 1990s. At the time of the adoption of this official state name [Belarus], there was extensive discussion as to the etymological link with the Russian name 'Belorussia'. The difference is in any case sufficiently great to consider there was never a Soviet Socialist Republic of Belarus, but a Belorussian SSR. Hence I use the name Belorussia for the Soviet republic, but Belarus for the post-Soviet state (and Belorussians up to 1991 but Belarusans afterwards)» [Smith 2013: xvi]. В других языках изменения не произошли вовсе или не распространилось за пределы официального дискурса. Так, например, в белорусских учебниках французского языка страна называется *le Bélarus* (в мужском роде), однако это название совершенно не используется носителями языка: на сайте посольства Франции в Минске сообщается, что несмотря на то, что ООН предлагает употреблять предложенный белорусским правительством вариант *Bélarus*, ряд французских институций (Национальная комиссия по топонимике, Министерство иностранных дел, Французская академия и др.) рекомендуют пользоваться словом *Biélorussie*[2]. В работе [Русакович 2014: 371], посвящённой названиям Белоруссии в немецком языке, сообщается, что «во второй половине XX века в научной литературе ФРГ и ГДР [наряду с *Weißrussland*] применялась также и транскрипция *Belorussland*», вышедшая в начале 1990-х из употребления, и сейчас встречаются как варианты *Weissrussland*, так и *Belarus*. В 2009 году Шведская академия решила употреблять в новом издании своего словаря прилагательное *belarusisk* [Лашкевич 2010], однако на награждении писательницы Светланы Алексиевич Нобелевской премией в 2015 году её гражданство обозначалось прилагательным *vitrysk*. Обзоры названий Белоруссии в разных языках можно найти в публицистических статьях [Лашкевич 2010], [Вячорка 2015a], [Вячорка 2015b].

Однако более интересно то, что происходило с вариативностью именно в русском языке, как общем языке Белоруссии и России — и в связи с этим в более конфликтогенной атмосфере.

Официальный русский перевод Конституции, принятой спустя 3 года после Закона о названии БССР — 15 марта 1994 года, начинался следующими словами:

---

[2] http://www.ambafrance-by.org/2-Toponymie-Belarus-ou-Bielorussie

«Мы, народ Республики Беларусь (Беларуси), исходя из ответственности за настоящее и будущее Беларуси, <…>»[3].

А в 1995 году русский язык стал государственным в Белоруссии, тем самым будто бы юридически закрепив наличие варианта *Беларусь* в русской литературной норме.

У нас нет сведений о том, насколько быстро топоним *Беларусь* закрепился в русскоязычном дискурсе белорусов, однако можно проследить его экспансию в прессе. Так, например, в главной государственной газете «Советская Белоруссия» (сейчас «СБ — Беларусь сегодня»), издающейся на русском языке, вариант *Беларусь* стал употребляться сразу же после принятия вышеупомянутого закона. Первое время, однако, в статьях могли встречаться оба топонима: старый преимущественно в тех случаях, когда речь шла о советских наименованиях:

(1) *Верховный Совет <…> делегировал в состав вновь образуемого Верховного Совета СССР представителей* ***Белоруссии*** *из числа народных депутатов СССР и народных депутатов* ***Республики Беларусь*** (СБ, 4 октября 1991 г.)

Кроме того, в публикациях 1991–1992 годов (в первую очередь, в письмах читателей) также могли изредка встречаться одновременно оба без семантически и прагматически мотивированных причин. Например, в статье под заголовком «**Беларусь** — не выдумка» в номере от 2 октября 1991 г. встречаются фразы «<…> поссорить **Белоруссию** и другие республики „без истории“ между собой и Францией» и «В последнее время мне доводится бывать в самых разных уголках **Белоруссии**»; ср. также следующий пример — два соседних предложения из письма читателя:

(2) *Мы хотим жить в свободной* ***Белоруссии****, <…>. Несмотря на все трудности, на экономическую разруху, у нас в* ***Беларуси*** *обстановка все-таки более стабильная.* (СБ, 28 апреля 1992 г.)

Старый топоним *Белоруссия* в 1991 и первой половине 1992 года встречается достаточно редко, но преимущественно без дополнительных коннотаций, со второй половины редкие использования этого топонима встречаются в письмах антинационалистически настроенных людей. Ср. фрагменты из двух писем, опубликованных в рубрике «Мнения: на разных полюсах»:

(3а) *Народ* ***Белоруссии*** *в основном — русскоязычный. <…> Сто пятьдесят лет русский язык был в* ***Белоруссии*** *государственным и должен им остаться. Теперь националисты, придя к власти, стремятся заменить русский язык белорусским.*

---

[3] https://ru.wikisource.org/wiki/Конституция_Республики_Беларусь/Первоначальная_редакция

(3b) *Вызывает удивление, что некоторые русские, проживающие в* ***Беларуси****, да и многие белорусы, требуют, чтобы обучение в школах, вузах, передачи по радио и телевидению велись на русском языке. <…> Неужели они не понимают, что в суверенной* ***Беларуси*** *русский язык такой же иностранный, как английский <…>.* (СБ, 7 мая 1992 г.)

Наконец, достаточно красноречив следующий вопрос от читателя:

(4) *Почему до сих пор ваша газета не изменила название? Ведь* ***Белоруссии*** *уже нет, тем более Советской.* (СБ, 30 ноября 1992 г.)

Примерно с 1993 года топоним *Белоруссия* в газете не встречается за исключением цитат из российских изданий. Таким образом, можно предположить, что к 1992–1993 году в русском языке Белоруссии вариант *Беларусь* закрепился достаточно крепко, постепенно вытесняя вариант *Белоруссия*.

3. Выбор топонима для Белоруссии в русском языке России

Несмотря на заявления белорусских политиков, в русском языке России основным остался вариант *Белоруссия*, который используется в большинстве СМИ и в обычной речи; в официально-деловых отношениях употребляются оба топонима. То же самое характерно и для других вариантов постсоветских топонимов — *Кыргызстан*, *Турмкенистан*, *Молдова*, *в/на Украине*, хотя и несколько в меньшей степени (кроме Украины). Подобное неполиткорректное с точки зрения представителей вышеназванных государств следование традиции в сочетании с выходом вариативности такого рода в интернет-пространство, где регулярно сталкиваются носители противоборствующих норм, привёл к значительному росту числа конфликтов в социальных медиа.

Случаи конфликтов в комментариях, где конфликтная коммуникация вызвана не содержанием исходного текста, а тем или иным употреблённым в ним словом, нередки. Например, в [Сомин 2015] описываются случаи, когда метаязыковые дискуссии в комментариях каузируются словом, относящимся не к тому стандарту белорусского языка, который ближе комментатору, или же вообще тем, что новостная статья написана на белорусском языке; во многих заметках в [Левонтина 2016] приводятся комментарии в социальных медиа, посвящённые отдельным словам из исходного текста (например, в [Левонтина 2016:272] таковым является слово *опа* из твита Ирины Родниной). Если же говорить только о конфликтах, связанных с топонимами, то важно отметить, что количество и мощность конфликтов *Беларусь/Белоруссия* (по крайней мере, в социальной сети «ВКонтакте») в десятки раз превышает аналогичные конфликты для других стран (не считая Украину, где, во-первых, актуальность и «горячесть» темы вызвана экстралингвистическими причинами, а во-вторых, поисковый механизм «ВКонтакте» не позволяет искать словосочетания).

Чтобы понять структуру этих конфликтов, для начала следует понять позиции сторон.

Очевидно, что в большинстве российских СМИ используется вариант *Белоруссия*. Для того же, чтобы узнать, как устроено распределение в речи обычных людей, мы провели следующий опрос-эксперимент, в котором приняли участие 574 человека. Участникам предлагалось определить десять государств, их язык и название жителей по флагу, столице и численности населения; одним из них была Белоруссия.

Для обработки результатов были исключены ответы всех тех, кто не из России. При подсчётах мы также исключили неверные (то есть не отмеченные в справочниках) варианты *Бел**а**руссия* (1,3%) и *Бел**о**русь* (5,5%), причины появления которых, впрочем, достаточно очевидны, фактические ошибки (например, *Украина*) и двойные ответы *Белоруссия/Беларусь* (1,5%). После всех сокращений в выборке осталось 418 ответов: 137 *Беларусь* (32,8% — из них 3,1% *Республика Беларусь*) и 281 *Белоруссия* (67,2%, из них 4,3% *Белору**с**ия*). Интересно, что для *Голландии/Нидерландов* распределение оказалось количественно практически таким же, но зеркально отражённым с точки зрения официальности топонима: 30,3% *Голландия* и 69,7% *Нидерланды*.

Можно предположить, что при подобных условиях опроса количество вариантов *Беларусь* — треть респондентов — выше, чем было бы, если бы у нас была возможность проанализировать спонтанную речь, так как в данном случае, во-первых, у участников опроса значительно больше степень рефлексии, чем в нейтральных условиях устной речи, а во-вторых, ситуация воспринимается как более формальная, что может каузировать употребление официального названия. Отметим, впрочем, что в быстрой речи половина падежных форм слов *Беларусь* и *Белоруссия* не различаются на слух.

Важно отметить, что для этнонимов и прилагательных распределение совершенно другое: 369 *бел**о**рус* (88,3%) и только 49 *бел**а**рус* (11,7%); 389 *бел**о**ру(с)ский* (93,1%) и только 29 *бел**а**ру(с)ский* (6,9%).

4. Выбор топонима для Белоруссии в русском языке Белоруссии

Как было показано во втором разделе, в русском языке Белоруссии используется исключительно вариант *Беларусь*. Это подтвердил и наш опрос, в котором оказалось 39 респондентов-белорусов (один человек всё же написал *Беларусь, Белоруссия*). При этом распределение для этнонимов и прилагательных напоминает аналогичное у россиян, хотя и с большим тяготением к варианту через *а*: 26 *бел**о**русы* (66,7%) и 12 *бел**а**русы* (31,8%), а в прилагательных — 32 *бел**о**русский* (84,2%) и только 6 *бел**а**рус(с)кий* (14,4%).

Таким образом, по-видимому, не только у россиян, но даже и у белорусов словарная норма оказывается зафиксированной в сознании гораздо сильнее, чем имена собственные, которые в традициях русской лексикографии не принято включать в словари. При этом этнонимы занимают промежуточное место

между топонимами как именами собственными и прилагательными, образованными от них, как именами нарицательными. Любопытно, впрочем, что в последнее время в белорусской русскоязычной публицистике стал всё чаще появляться вариант *беларус* и прилагательное *беларусский* или даже *беларуский*, например, в книгах под редакцией публициста Анатолия Тараса, ср. «История имперских отношений. Беларусы и русские. 1772–1991» (Минск, 2008) или «Деды. Дайджест публикаций о беларуской истории. Специальный выпуск. Выпуски 1–5» (Минск, 2013). Наиболее же ранняя публичная фиксация этой неоднозначности имеется всё в той же «Советской Белоруссии» в письме от читательницы:

(5) *Уважаемый парламент, уважаемые филологи. Будьте добры и ответьте на такой наивный вопрос: как правильно написать в метрике новорождённого его национальность — «белорус» или «беларус».*
(СБ, 2 декабря 1992 г.)

***

Для понимания сути конфликтов важно понять причину этих возмущений. Для этого мы провели опрос в интернете, а также двенадцать пилотных устных интервью в Минске. В силу небольшого количества респондентов в интервью полученными данными можно пользоваться лишь с оговорками, однако нейтральное или позитивное отношение к варианту *Белоруссия* наблюдалось только у людей старше 40 лет.

Результаты онлайн-опроса, однако, несколько отличаются. В нём приняло участие 74 человека в возрасте от 18 до 75 лет (опрос предлагался лишь белорусам, однако в число респондентов вошли три человека, живущих всю жизнь в России, которые были исключены из выборки; несколько человек, проживших в России от 5 лет и больше, были всё же оставлены). В опросе, кроме сбора социолингвистических данных, предлагалось ответить на вопрос про отношение к варианту *Белоруссия*, указать возраст, в котором пришло осознание наличия двух норм, выбрать вариант этнонима и указать своё отношение к другому, а также сообщить, станет ли респондент исправлять собеседника, использующего неверный вариант, и описать аргументы, к которым он будет при этом апеллировать.

В результате 52 из 71 опрошенного (73,2%) выразили более или менее негативное отношение — от ненависти до лёгкого раздражения — к варианту *Белоруссия*, 15 (21,1%) — нейтральное, и лишь четверо (5,6%) описали вариант *Белоруссия* как единственно правильный. Следует, однако, отметить, что средний возраст составил 26,4 года (а если исключить 75-летнего информанта, то 25,7), медианное значение — 24, что в некотором роде объясняет результаты: сознательная жизнь абсолютного большинства опрошенных прошла уже в независимой Республике Беларусь. Возрастом начала рефлексии большинство опрошенных указывают 14–16 лет, часть из них прямо указывают на роль в этом школы, гораздо реже — начало учёбы в университете.

Наиболее интересен же следующий результат: 34 человека (47,9 %) указывают этноним *бел**а**рус* как единственно верный или более предпочтительный из пары и только 23 (32,4 %) считают единственно верным или более предпочтительным вариантом зафиксированное в словарях *белорус* (в их число также входят люди, предпочитающие вариант *Белоруссия*, то есть только среди сторонников *Беларуси* этот процент ещё меньше). При этом многие из тех, кто отдаёт предпочтение написанию через *о*, утверждают, что считают более правильным (логичным, подходящим и т. п. — воспроизводящим орфографию топонима) вариант через *а*, но вынуждены следовать словарям. По-видимому, можно говорить об изменении нормы на наших глазах: вполне возможно, что, если исключить экстралингвистические факторы, через один или несколько десятков лет написание *беларус* (и, вероятно, *беларус(с)кий*) может вытеснить в белорусском регионолекте русского языка словарный вариант. Этот факт, насколько нам известно, на данный момент не отражён ни в одной академической работе (наиболее близко — [Hroch 2010:280]: «<…> to use the Belarusian language (preferred today to Belorussian <…>»).

Анализируя полученные данные, не следует, однако, забывать, что с большей вероятностью участие в подобном опросе примет человек, который интересуется данной темой и рефлексирует по этому поводу; соответственно, можно ожидать некоторого перекоса в сторону некодифицированных вариантов. И действительно, анализ спонтанного употребления конфликтно немаркированных лексем в ситуациях нейтральной (неконфликтной) коммуникации даёт иные результаты. Согласно данным, предоставленным В. И. Беликовым[4], в белорусских блогах за 2007–2008 годы наблюдалась следующая статистика употребления наименований учителей белорусского языка:

	блоги (2007–2008)	ВКонтакте (всё время)
*бел**о**русичка / бел**о**русица:*	4/45	29/72
*бел**о**руссичка / бел**о**руссица*:	2/14	5/8
*бел**а**русичка / бел**а**русица*:	3/20	10/45
*бел**а**руссичка / бел**а**руссица*:	0/2	1/1

Здесь отношение написания через *о* к написанию через *а* составляет 65/25, т. е. 2,6 к 1 (отдельного внимания также заслуживают 18 случаев гиперкорректного *сс*). Аналогичные результаты показывает и статистика в социальной сети «ВКонтакте» за все годы (с последующей ручной обработкой для исключения употреблений в другом значении, как, например, *Белорусичка* как ироничное название страны или *белорусиц* как презрительное название национальности) — 57/114, т. е. 2 к 1. Кроме того, в подобных контекстах часть написаний через *а* может быть не показателем принципиальной позиции пишущего, а обычной орфографической ошибкой. Таким образом, с учётом соотношения употреблений двух вариантов написания в сфере неконфликтного

[4] Мы выражаем искреннюю признательность В. И. Беликову за статистику и саму идею обратить внимание на подобные употребления, а также за другие ценные комментарии.

(и неотрефлексированного) общения скорый переход к официальному признанию написания через *а* не кажется столь вероятным.

Что же касается последнего вопроса — об исправлении собеседника, то из 52 респондентов, относящихся к варианту *Белоруссия* негативно, 36 человек (независимо от пола) утверждают, что будут поправлять употребившего данный вариант собеседника, ещё у шестерых это зависит от ситуации общения или собеседника (итого 80,7% из «негативной» группы, не включая двух человек из «нейтральной»). Таким образом, если считать нашу выборку репрезентативным срезом белорусов в интернете, 53,5% (38 из 71) готовы вступить в спор с собеседником, чтобы научить того правильному названию страны. Это объясняет частоту и мощность споров на данную тему в социальных медиа.

В чём причина столь ярой нелюбви белорусов к варианту *Белоруссия*? Большинство опрошенных ссылаются на то, что они воспринимают слово *Белоруссия* как пережиток прошлого — название советской республики, и этот вариант противоречит официальному названию, указанному в русскоязычной конституции. Многие характеризуют это слово как уничижительное или пренебрежительное; оно не только фонетически близко к названию России (что в том числе запутывает иностранцев), но и по своему морфологическому и фонетическому составу якобы указывает на зависимое, периферийное относительно России положение государства, будто бы российский регион. Люди же, употребляющие этот вариант, для части информантов просто неграмотные, ещё для части — невежливые, необоснованно пренебрегающие мнением граждан этого государства. Несколько человек указали на требование транслитерировать название в соответствии с белорусским звучанием, и, наконец, ещё несколько сравнили русский язык России и Белоруссии с американским и британским английским, указывая на то, что для языка нормально быть полицентричным и иметь региональные нормы. Если же говорить об аргументах, которые будут приводиться в гипотетических спорах, то в абсолютных лидерах ссылки на конституцию, объяснение, что это название советской республики, и просто указание на нелюбовь белорусов к этому слову (апелляция к вежливости).

Несмотря на столь явные выводы, насколько нам известно, лишь в двух работах упоминается негативное восприятие белорусами названия *Белоруссия*. Во-первых, это публицистическая статья [Заварин 2015:24], где разбираются 10 мифов об этой стране: в мифе №6 «Беларусь — это Белоруссия» сообщается, что «белорусов, которые не сильно скучают по советскому прошлому, сильно раздражает название „Белоруссия“. Если не хотите их обидеть, следуйте правилам русского языка». Во-вторых, глава в учебнике [Sloboda 2011:396], отмечающая, как и респонденты нашего опроса, следующее: «The more traditional English names „Belorussia“ and „Byelorussia“ are derived from the Russian name of Belarus (*Belorussia*). They thus symbolize the former dominance of Russia over Belarus and are considered politically incorrect by the Belarusians today».

Важно подчеркнуть, что выбор номинации *Беларусь* в дискурсе белорусов, равно как и — в большинстве случаев — употребление топонима *Белоруссия* россиянами является немаркированным, не свидетельствующим о каких бы то ни было националистических в первом случае или, во втором случае,

«империалистских» установках говорящего. В обратную же сторону оппозиция привативна: вариант *Белоруссия* в речи белорусов, особенно старшего поколения, скорее, нейтрален, тогда как употребление топонима *Беларусь* россиянином с наибольшей вероятностью окажется маркированным и указывающим на высокую степень осознанности выбора номинации.

5. Структура конфликтов *Беларусь* vs. *Белоруссия* в социальных медиа

Теперь, когда мы разобрались с причинами конфликтов, можно описать их структуру. Возможность их появления обуславливается общностью медиасферы России и Белоруссии: многие белорусы читают новости на российских ресурсах (например, сайт rbc.ru находится на 42-м месте в топе сайтов по посещаемости из Белоруссии), а социальная сеть «ВКонтакте» является в принципе наиболее посещаемым сайтом в Белоруссии[5].

Конфликты, происходящие в интернете, в большинстве случаев начинаются с возмущения белоруса написанием названия его страны (и это логично: россияне практически не бывают на белорусских сайтах), хотя бывают и исключения, ср.:

(6) *Извините, не могу молчать. Когда малограмотный пацан не знает, что по-русские пишется «Белоруссия», это понятно. И простительно — что взять с него. Но когда лингвист…*
В русском языке нет слов «Пари», «Ландон» и «Беларусь». В русском есть слова «Париж», «Лондон» и «Белоруссия». Пожалуйста, заучите. Географические названия приводятся в соответствии с исторической традицией языка перевода, а не в соответствии с фонетикой языка-исходника. (комментарий в блоге, 26 мая 2013)

Далее, в ответ на возмущение белоруса российский пользователь сообщает, что по нормам русского языка правильно — *Белоруссия*. Если спор продолжается, то стороны в основном используют следующие аргументы:

сторонники *Белоруссии:*

1) названия других государств и городов не транслитерируются на русский язык: *Дойчланд, Франс, Рома*;
2) вариант *Беларусь* не соответствует правилам русского языка (не бывает соединительной гласной *а*), это белорусское слово;
3) изменения названий постсоветских стран не нашли отражения в русском языке, придерживающемся традиционных норм (*Молдавия*, *Киргизия* и т. п.), одно государство не может указать другому, как его называть — особенно если речь идёт о национальном языке этого

[5] http://www.alexa.com/topsites/countries/BY

государства (т. е. в данном случае русский язык рассматривается как собственность России);
4) отказ использовать вариант, закреплённый в конституции Республики Беларусь в качестве официального и, соответственно, не обязательного в обычной речи;
5) согласие с обоими вариантами как допустимыми;
6) шуточные аргументы (в основном «Беларусь — это трактор, а страна — Белоруссия»),

а также несколько конкретных ссылок:
7) на письмо ИРЯ РАН по запросу Яндекса[6], которое указывает, что вариант *Белоруссия* используется в бытовой сфере общения и «носители русского языка вправе использовать его в соответствующей ситуации» (на это письмо также любят ссылаться сторонники *Беларуси*, т. к. в нём говорится, что «оба наименования — *Белоруссия* и *Беларусь* имеют право на существование и употребление в современном русском языке»);
8) на фотографию президента А. Г. Лукашенко, сидящего на саммите с табличкой *Республика Белоруссия*;
9) на статью в Википедии «Именование белорусского государства на русском языке»;

сторонники *Беларуси*:
1) русский — государственный язык Беларуси, название закреплено в русскоязычной конституции и указано в паспортах (а также в законе 1991 года, цитируемом в начале статьи), а те страны, названия которых не транслитерируются (Германия, Франция и т. п.), не имеют русского в качестве государственного;
2) по закону 1991 года название должно транслитерироваться на все языки мира в соответствии с белорусским произношением;
3) слово *Беларусь* используется русскоязычными жителями этой страны, а статус русского языка как государственного легитимизирует право этих людей влиять на его нормы;
4) *Белоруссия* — название советской республики, *Беларусь* — название независимого государства, и страны с названием *Белоруссия* не существует (так же, как Бирма стала Мьянмой, Берег Слоновой Кости — Кот-д'Ивуаром);
5) ссылки на российские (полу)официальные документы: письмо ИРЯ РАН, Общероссийский классификатор стран мира, сайт Института географии РАН;
6) ссылка на вышеупомянутую статью в Википедии;
7) шуточные комментарии (в основном, указания на существование страны *Россисии*, если существует *Белоруссия*).

---

[6] http://bygirl.net/wp-content/uploads/2012/03/belarusvsbelorussia.jpg

Так, например, в следующем примере со стороны защитников варианта *Беларусь* используются аргументы № 1, 4, 5, а со стороны защитника *Белоруссии* — № 1, 2, 3, 4 и 5:

(7) *Максим: Константин, нет такой страны, как Белоруссия. Есть Беларусь. Не трудно запомнить. <…>*
Деградат нация: Максим, нет такой страны, как Германия, Англия и Russia. Есть только Deutschland, England и Россия.
Максим: Деградат нация, хм. Меня тут на днях одна щепетильная особа доставала тем, что я сказал Белоруссия. Вот теперь я тоже проповедую знания)) Не более))
Глеб: Деградат нация, есть общероссийский классификатор стран мира, там нет ни Англии ни Белоруссии. Тем более, название «Республика Беларусь» закреплено официально именно на русском языке. Так что уви.
Деградат нация: Глеб, ровно как и Белоруссия, например. Оба названия употребляются, и их употребление не является ошибкой.
Глеб: Деградат нация, Наименование страны Краткое БЕЛАРУСЬ . Полное Республика Беларусь. Белорусия устарела в 1991 году.
Александр: Глеб, в русском языке допустимо слово Белоруссия, беларусы по правилам русского языка пишутся через о, и как понять устарело? Да, до 91 года Белорусская ССР, коротко ее называли Белоруссией, в простонародье и до сих пор называют.
Глеб: Александр, так и понять. Новая страна, новое название. «Да потому что нет больше вашей Белоруссии, ха-ха-ха-ха»
Александр: Глеб, и? У каждого языка свои особенности правописания и произношения названий других народов и стран, выше об этом уже говорилось, приводя в пример Германию, Китай и тд. И никто не вправе указывать им.
Глеб: Александр, это не особенность белорусского, это именно закреплено в конституции на русском языке.
Александр: Глеб, причем тут конституция? Есть русские словари, там есть слова Белоруссия, белорусы и тд. Мы не научно\официальным деловым стилем разговариваем, чтобы в конституцию смотреть.

Достаточно интересно, что некоторые причины недовольства, отмеченные в проведённом нами опросе, не приводятся в качестве аргументов (и это понятно: ассоциация названия *Белоруссия* с зависимостью от России — плохой аргумент в споре).

6. Заключение

Качественный анализ конфликтных дискуссий в социальных медиа позволил нам выделить основные тезисы, используемые сторонниками вариантов *Беларусь* и *Белоруссия* при спорах в комментариях. Аргументы сторонников *Беларуси* не полностью повторяют реальные причины их недовольства вариантом

Белоруссия. Дополнительные исследования (эксперименты, опросы, интервью, работа с газетами начала 1990-х) показывают причины возникновения этих разногласий и объясняют суть некоторых аргументов из споров в комментариях. Кроме того, отмечено, что несмотря на словарную фиксацию, вариативность затрагивает также и прилагательное и этноним; при этом можно предположить, что вариант национальности *беларус* (а также и образованное от него прилагательное) в русском языке Белоруссии в будущем может вытеснить словарную норму *белорус*, хотя результаты количественного анализа социальных медиа — подсчёта употреблений слов, однокоренных с обоими вариантами топонима, снижают категоричность этой гипотезы.

В дальнейшем представляет интерес, во-первых, сравнение структуры конфликта *Белоруссия/Беларусь* с другими лингвополитическими конфликтами (*Молдавия/Молдова*, *на Украине/в Украине* и т. п.), а во-вторых, рассмотрение сценариев развития лингвополитических споров на фоне других конфликтов, индуцированных языком, в частности, связанных с использованием регионализмов, известных не всем участникам коммуникации, и слов-раздражителей, в том числе из языков субкультур (*кушать*, *печеньки, пузожитель, годовасик* и т. п.).

Литература

1. *Вячорка, В.* 2015a. Як нас заве сьвет — «Беларашэн» ці Belarus(i)an? http://www.svaboda.org/content/article/27189235.html (проверено 17.02.2016)
2. *Вячорка, В.* 2015b. Ці абароніць Belarus Сьвятлана Алексіевіч? http://www.svaboda.org/content/sviatlana-aleksievich-belarus/27310245.html (проверено 17.02.2016)
3. *Заварин, С.* 2015. Lonely republic: Беларусь — территория легенд // Профиль № 891 (2). С. 20–25.
4. *Кронгауз, М. А.* 2013. Кто отвечает за русский язык. // Русский язык зарубежья. СПб: Златоуст. С. 4–14.
5. *Лашкевич, К.* 2010. От «Белоруссии» к «Беларуси»: процесс пошел. http://news.tut.by/politics/157124.html (проверено 17.02.2016)
6. *Левонтина, И. Б.* 2016. О чём речь. М.: Corpus.
7. *Русакович, А. В. 2014.* Эволюция транскрипции названия «Беларусь» в германской научной литературе в 1990–2000-е гг. // Беларусь в современном мире. Материалы XIII Международной конференции, посвященной 93-летию образования Белорусского государственного университета. Минск: БГУ. С. 370–372.
8. *Сомин, А. А.* 2015. Языковая рефлексия в современной Беларуси сквозь призму комментариев в интернет-СМИ // Вестник РГГУ. Серия «История. Филология. Культурология. Востоковедение» / Московский лингвистический журнал. Том 17 (1). С. 62–86.
9. *Hroch, M.* 2010. The Slavic World, in Fishman, J. A., García, O. (ed.) Handbook of language and ethnic identity: Disciplinary and regional perspectives. Volume 1. Second edition. Oxford university press. Pp. 269–285.

10. *Sloboda, M.* 2010. Belarusian, in Fishman, J. A., García, O. (ed.) Handbook of language and ethnic identity: The success-failure continuum in language and ethnic identity efforts. Volume 2. Oxford University Press. Pp. 381–398.
11. *Smith J.* 2013. Red nations. The nationalities experience in and after the USSR. Cambridge University Press.

References

1. *Hroch M.* (2010), The Slavic World, in Handbook of language and ethnic identity: Disciplinary and regional perspectives. Volume 1. Second edition, ed. by Fishman, J. A., García, O. Oxford university press. Pp. 269–285.
2. *Krongauz M. A.* (2013), Who is responsible for the Russian language [Kto otvechaet za russkij jazyk], in The Russian language of foreign countries [Russkij jazyk zarubezhja], Zlatoust, Saint-Petersburg. Pp. 4–14.
3. *Lashkevich K.* (2010), From „Belorussia" to „Belarus": the process has started [Ot "Belorussii" k "Belarusi": protsess poshël], available at http://news.tut.by/politics/157124.html.
4. *Levontina I. B.* (2016), What is it all about [O chëm rech'], Corpus, Moscow.
5. *Rusakovich A. V.* (2014), Evolution of the transcription of the name "Belarus" in German science literature in 1990–2000 [Èvolutsija transkriptsii nazvanija "Belarus'" v germanskoj nauchnoj literature v 1990–2000-je gg.], Belarus in the modern world: Proceedings of the XIII International conference dedicated to 93 years from the foundation of the Belarusian State Univercity [Belarus' v sovremennom mire. Materialy XIII Mezhdunarodnoj konferentsii, posvjashchënnoj 93-letiju obrazovanija Belorusskogo gosudarstvennogo universiteta], Minsk, pp. 370–372.
6. *Sloboda M.* (2010), Belarusian, in Handbook of language and ethnic identity: The success-failure continuum in language and ethnic identity efforts. Volume 2. ed. by Fishman, J. A., García, O. Oxford University Press. Pp. 381–398.
7. *Smith J.* (2013), Red nations. The nationalities experience in and after the USSR. Cambridge University Press.
8. *Somin A. A.* (2015), Attitudes to the Belarusian language through the prism of comments in the online media [Jazykovaja refleksija v sovremennoj Belarusi skvoz' prizmu kommentariev v internet-SMI], RGGU Buletin. Series "History. Philology. Cultural studies. Orientology" / Moscow Journal of Linguistics [Vestnik RGGU. Serija „Istorija. Filologija. Kul'turologija. Vostokovedenie" / Moskovskij lingvisticheskij zhurnal], Vol. 17 (1), pp. 62–86.
9. *Viačorka V.* (2015a), How does the world call us — „Bielarašen" or Belarus(i)an? [Jak nas zavie świet — «Bielarašen» ci Belarus(i)an?] available at http://www.svaboda.org/content/article/27189235.html
10. *Viačorka V.* (2015b), Will Svetlana Alexievich protect „Belarus"? [Ci abaronić „Belarus" Śviatlana Alieksijevič?], available at http://www.svaboda.org/content/sviatlana-aleksievich-belarus/27310245.html
11. *Zavarin S.* (2015), Lonely republic: Belarus — the territory of legends [Lonely republic: Belarus' — territorija legend], Profile [Profil'], Vol. 891 (2). Pp. 20–25.

SPELLRUEVAL: THE FIRST COMPETITION ON AUTOMATIC SPELLING CORRECTION FOR RUSSIAN

Sorokin A. A. (alexey.sorokin@list.ru)[1,3,4],
Baytin A. V. (baytin@yandex-team.ru)[2],
Galinskaya I. E. (galinskaya@yandex-team.ru)[2],
Rykunova E. D. (alenarykunova@gmail.com)[3],
Shavrina T. O. (rybolos@gmail.com)[1,4]

[1]Lomonosov Moscow State University, Moscow, Russia
[2]Yandex, Moscow, Russia
[3]Moscow Institute of Physics and Technology, Dolgoprudny, Russia
[4]General Internet Corpora of Russian, Moscow, Russia

This paper reports on the first competition on automatic spelling correction for Russian language—SpellRuEval—held within the framework of "Dialogue Evaluation". The competition aims to bring together groups of Russian academic researchers and IT-companies in order to gain and exchange the experience in automatic spelling correction, especially concentrating on social media texts. The data for the competition was taken from Russian segment of Live Journal.

7 teams took part in the competition, the best results were achieved by the model using edit distance and phonetic similarity for candidate search and n-gram language model for their reranking. We discuss in details the algorithms used by the teams, as well as the methodology of evaluation for automatic spelling correction.

Key words: spelling correction, automatic spelling correction, language of social media, automatic methods for processing Russian

SPELLRUEVAL: THE FIRST COMPETITION ON AUTOMATIC SPELLING CORRECTION FOR RUSSIAN

Сорокин А. А. (alexey.sorokin@list.ru)[1,3,4],
Байтин А. В. (baytin@yandex-team.ru)[2],
Галинская И. Е. (galinskaya@yandex-team.ru)[2],
Рыкунова Е. Д. (alenarykunova@gmail.com)[3],
Шаврина Т. О. (rybolos@gmail.com)[1,4]

[1]МГУ им. М. В. Ломоносова, Москва, Россия;
[2]Яндекс, Москва, Россия;
[3]МФТИ, Долгопрудный, Россия;
[4]ГИКРЯ, Москва, Россия

В этой статье обсуждается первое соревнование по автоматическому исправлению опечаток на материале русского языка, SpellRuEval, прошедшее в рамках проекта “Dialogue Evaluation”. Целью соревнования является сравнение разнообразных методов и подходов, применяемых для и исправления опечаток, а также обмен опытом между научными коллективами и IT-компаниями, имеющими свои успешные разработки в этой области. Соревнование проводилось на материале блогов Живого Журнала.

В данной статье подробно разбираются результаты, полученные от 7 коллективов, участвовавших в соревновании, сравниваются подходы, применённые участниками соревнования. Наилучшие результаты были достигнуты моделью, использовавшей редакционное расстояние для поиска кандидатов и комбинацию взвешенного редакционного расстояния и n-граммной языковой модели для отбора наилучшего исправления. Также в статье подробно обсуждается методика оценки качества автоматического исправления опечаток.

Ключевые слова: исправление орфографии, автоматическое исправление орфографии, язык социальных медиа, исправление опечаток

1. Introduction

SpellRuEval is the first competition aimed to make a framework for evaluation of automatic spelling correction systems for Russian and cooperation and experience exchange of scientific groups. Today, when huge amounts of data are collected from Russian internet resources (e.g. Yandex Blogs, RuTenTen Corpora, Russian Araneum Corpora and GICR), automatic processing of this data is an unavoidable problem [Manning 2011]—misspells widely hinder morphological, syntactic and semantic parsing of the texts. By the estimation of [Baytin, 2008], 15% of all the queries in Yandex have at least 1 error, and by the data of [Shavrina, Sorokin, 2015] nearly 8% of the out-of-vocabulary words (further “OOV”) are typos. Moreover, for some data sources the percentage of typos may reach 40% (Private communication, GICR).

Hence, there emerges a bulk of actual challenges for NLP-researches: which error detection model for Russian internet text is the best—dictionary look-up or rule-based? Which models are the best for isolated error-correction and which are better for context errors? How to raise the quality of real-word error detection and correction? Is there any dependency between dictionary size and recall of spelling detection? Which algorithms of machine learning give the best results for spelling correction on social media texts? All these problems we have faced during the preparation of the competition procedure and the analysis of the results.

1.1. A brief history of automatic spelling correction

Automatic spelling correction is one of the oldest problems of computational linguistics. The first theoretical works appeared already in the 60-s [Damerau, 1964]. The initial approach used edit (Levenshtein) distance [Levenshtein, 1965] to search

for potential corrections of mistyped words. With the appearance of modern spellcheckers [McIlroy. 1982] in the early 80-s, the problem of spelling correction became a highly practical one. The most important papers appeared on the dawn of modern NLP era include [Kernighan et al., 1988], [Mays et al., 1991] and [Kukich, 1992], which is in excellent review of early approaches in automatic spelling correction. Further work in spelling correction was developed in two main directions: the works of the first category mainly addressed the problem of effective candidate search, which is a non-trivial problem for the languages with well-developed morphology [Oflazer, 1996], [Schulz, Mihov, 2002]. This branch also includes the research on learning adequate distance measure between the typo and the correction [Ristad, Yanilos, 1998], [Kernighan et al., 1990], [Brill, Moore, 2000], [Toutanova et al., 2002]. Other researchers mainly addressed the problem of using context when selecting the correct candidate for spelling correction. The most important works here include [Golding, Schabes, 1996], [Golding, Roth, 1999], [Hirst, Budanitsky, 2005], [Cucerzan, Brill, 2004].

The problem of automatic spelling correction includes several important subtasks. The first is to detect whether a word has correct spelling and provide a list of candidates. As observed by many researchers, most of the time the correction can be obtained from the mistyped word by single letter deletion, insertion or substitution or by permutation of two adjacent characters [Kukich, 1992]. However, in many cases this procedure yields multiple candidate words and additional features should be taken into account to select the most proper one. This is especially a problem for agglutinative languages or languages with a high number of inflected forms since a single edit operation on a word often creates another form of the same word and morphology and syntax should be used to disambiguate between them. The so-called real-word errors (when a mistyped word is again in the dictionary) constitute the most difficult problem. Several researchers addressed it [Liu, Curran, 2006], [Carlson, Fette, 2007], [Pedler, Mitton, 2010], however, all the algorithms were tested on pre-defined confusion sets, such as 'adopt/adapt' and 'piece/peace', which makes rather problematic the application of their methods to real-word errors outside these sets.

Evaluation of spellchecking techniques presents another difficult challenge. Indeed, spelling correction is applied in different areas, mainly for Internet search and information retrieval [Ahmad, Kondrak, 2005], [Cucerzan, 2004], [Zhang, 2006], [Whitelaw, 2009] and in text editors, but also in second language acquisition [Flor, 2012] and grammar error correction [Rozovskaya, 2013]. The area obviously affects the character of typical spelling errors. Moreover, the effect of different features for spelling correction also highly depends from the application. Morphology and especially syntax give little advantage in case of search query correction, in this case the quality of the dictionary and gazetteer, as well as size of query collection used to train language and error models, is more important. In case of grammar error correction the situation is roughly the opposite. Most of spelling correction systems were tested on rather artificial or restricted datasets: the authors either asked the annotators to reprint the text without using 'backspace' and 'delete' keys [Whitelaw, 2009] or used Wikipedia [Schaback, 2007] or TOEFL essays collection [Flor, 2012]. Often the authors just randomly replaced a word by a potential misspelling, using some error model ([Carlson, Fette, 2007] etc.) Therefore it is not obvious that results obtained in one subarea could be successfully used in the other one.

2. Related Work

Most of spellchecking approaches were tested on English language, which is certainly not the most difficult for this task. First, a large collection of corpora is available for English and additional data could be easily collected from the Web. Second, English is very simple from the morphological point of view, therefore most of the problems concerning morphology or dictionary lookup even does not arise there. There are very few works for other languages with complex and diverse morphology, such as Turkish or Arabic ([Oflazer, 1996], [Mohit et al., 2014], [Rozovskaya et al., 2015]). The studies for Russian language include only [Baytin, 2008], [Panina et al., 2013] and [Sorokin and Shavrina, 2015], but all these works also address spelling correction problem in a rather restricted way.

2.1. First automatic spelling correction contests

In the field of automatic spell-checking for English two works can be considered as pioneer. These are Helping Our Own (HOO) Shared Tasks of 2011 and 2012 correspondingly [Dale et al., 2011] and [Dale et al., 2012]. Although the theme of the competition was set broader than just spelling correction (the main goal was to map and develop tools that can assist authors in the writing task and facilitate the processing of the typed texts), these competitions obviously exhibited the main problems of state-of-art methods and led to more specified workshops, such as Microsoft Spelling Alteration Workshop [Wang, 2011]. It was primarily concerned with correcting errors in search queries: participant systems were evaluated on the logs of Bing search engine. The close problem of grammatical error correction was the thematics of CoNLL 2013 Shared Task [Ng et al., 2013]. However, all these competitions were held for English Language. There were no such competition for Russian and even a freely available dataset of spelling errors, such as Birkbeck corpus for English [Mitton, 1986] did not exist. The primary purpose of SpellRuEval-2016 was to fill this gap and evaluate different approaches to automatic spelling error correction for such morphologically and syntactically complex language as Russian.

2.2. First and second QALB Shared Task on Automatic Text Correction for Arabic

The first competition to succeed on automatic text normalization and spelling correction, which was carried out on not English-based materials, was the first QALB Shared Task on Automatic Text Correction for Arabic [Mohit et al., 2014]. The competition united more than 18 systems and determined a baseline of 60–70% (Precision), which is quite a progress for such languages as Arabic. By this time, there was already held the second QALB Shared Task [Rozovskaya et al., 2015] with the improvement of the baseline up to 80% of Precision. Both of the competitions were based on the Qatar Arabic Language Bank, however, they focused on slightly different goals: if the

first QALB shared task was to correction of misspells, punctuation errors, extra spaces and normalization of the dialecticisms on the corpus of native speakers, the second one have added the corpora of L2-speakers in the training set, that shifted the researchers' attention to frequent mistakes made by learners of Arabic.

3. Procedure of SpellRuEval competition

3.1. Training and test data

In this section we describe the format of training and text data used in the competition. We used a Live Journal subcorpus of General Internet Corpora of Russia (GICR) [Belikov et al., 2013] to extract test sentences. We automatically selected about 10,000 sentences containing words not present in the dictionary. The sample was enriched by several hundred sentences containing real-word errors; these sentences were obtained from the same source corpus. Then we manually filtered these sentences to ensure that these sentences indeed contain typos, not rare proper names, slang or neologisms. About 5,000 remaining sentences were loaded to the annotation system. We asked the annotators to correct the typos in each sentence following a short instruction and submit the corrected sentence. If the annotator met a controversial case, supposed to be not covered by the instruction, he or she could also submit the commentary, explaining the difficulty.

The instruction contained the following items:

1. The annotator should correct:
   a) typos (*мнея → меня*),
   b) orthographic errors (*митель → метель*),
   c) cognitive errors (*компания → кампания*),
   d) intentional incorrect writing (*хоцца → хочется, ваще → вообще*),
   e) grammatical errors (agreement etc.) (*он видят → он видит*),
   f) errors in hyphen and space positioning (*както→как-то*),
   g) mixed usage of digits and letters in numerals (*2-ух →двух*),
   h) usage of digits instead of letters (*в4ера → вчера*).
2. The annotator should not correct
   a) foreign words including cyrillic (e.g Ukrainian or Belorussian),
   b) informal abbreviations (*прога → программа*)
   c) punctuation errors (all punctuation is omitted during the testing procedure—for more details, see chapter 3.2)
   d) capitalization errors (as capitalization is rather varied and informal in Russian social media texts, see also 3.2)
   e) non-distinction of "e" and "ë" letters

Most of the controversial moments in the annotation dealt with colloquial forms such as *ваще* for *вообще* and *щас* for *сейчас*. In most of the cases they can be freely replaced by corresponding formal forms without any change in meaning, except for

the expressive sentences like *«Ну ты ваще»* (1) or *«Да щас!»* (2), so in the latter cases there is no typo to correct. But obviously the border between these cases is very subtle so we deliberately decided to correct such colloquial forms in all the sentences.

Each of the 5,000 sentences was given to three annotators. Most of the annotators were the competition participants or students of linguistic and computer science departments. The annotation logs were automatically processed to select the sentences where all the three annotators gave the same answer and then manually filtered to avoid prevalence of several frequent typo patterns. Finally, about 2,400 mistyped sentences remained. The sample was extended by 1,600 correctly typed sentences obtained from the same corpora. The final sample of 4,000 sentences was randomly subdivided by two equal portions, each containing 2,000 sentences. The first half was given to the competitors as the development set. Such small size of the development set gave the participants no possibility to learn language model, however, they could use this sample to tune the parameters of their algorithm: e.g. the weighted Levenshtein distance used for candidate search or the weights of different features (error model, language model, morphology model etc.) in the final decision procedure. We also provided an evaluation script using which the participants could measure the performance of their systems on the development set. Since we have not provided any dictionary or corpora resources, the competitors were allowed to use arbitrary dictionary to search for candidates and arbitrary corpus, say, to fit the language model.

Since 2,000 sentences can be manually corrected in one or two days, they were randomly inserted into the sample of 100,000 sentences taken from the same corpus. The participants had no information about this fact and were asked to send the answers for the whole test sample. However, the correctness was evaluated only on the 2,000 sentences from the test set.

3.2. Evaluation and metrics

The proper selection of evaluation metric for the competition was not a trivial task. A common metric for Web search spelling correction is the fraction of correctly restored queries, its direct analogue is the percentage of correct sentences. However, it is uninformative for our task: this metric cannot show the difference in performance between a system with high recall which corrects all the typos but also a lot of correctly typed words, and a system which has high precision and corrects no sentences at all. This problem could be partially remedied by calculating the number of properly and improperly corrected sentences with typos, as well as the number of "false alarms" (improperly corrected sentences without typos), but this metric is also inadequate when sentences could contain several typos. For example, consider a sentence with two typos, the described evaluation algorithm cannot distinguish a sentence with only one typo corrected from a sentence with two typos corrected and properly and one correct word changed incorrectly.

Therefore we evaluate performance in terms of individual corrections, not the whole sentences. That raises the problem of sentence alignment: in the case of space or hyphen orthographic error one word in the source sentence may correspond

to multiple words in the correction, as well as many words in the source to a single one in the corrected sentence. We aligned the sentences using the following procedure:

1) First, the sentence was converted to lowercase and split by the space symbols.
2) The isolated punctuation marks were removed.
3) Since most of the punctuation symbols are not separated from the previous words, all non-alphabetic characters were deleted on both edges of each word.
4) Then the source sentence and its correction were aligned using the following algorithm:
    1. Longest common subsequence was extracted using standard dynamic programming algorithm. Words on the same position in the subsequence were aligned to each other.
    2. Each of the nonidentical groups between alignment points constructed on the previous step was aligned separately. We constructed a Levenshtein alignment between source and correction sides of the groups using standard edit distance with permutations, separating the words in groups by spaces. If an alignment point was located between the words both on the source and correction sides, then this point was added to the alignment.

Below we explain this algorithm on the sentence *«помоему, кто то из них то же ошипся»* (3) and its correction *по-моему, кто-то из них тоже ошибся*. After the removal of punctuation marks and first step of the alignment algorithm we obtain the following alignment groups:

(4)

помоему кто то	*по-моему кто-то*
из	*из*
них	*них*
то же ошипся	*тоже ошибся*

When processing the pair (*помоему кто то, по-моему кто-то*), we observe that an optimal alignment matches the groups *«помоему»* and *«по-моему»* to each other. Since both these subgroups end on word edges, we obtain additional alignment pairs

(5)

помоему	*по-моему*
кто то	*кто-то*

and the remaining part

из	*из*
них	*них*
то же	*тоже*
ошипся	*ошибся*.

After constructing such alignment for all the pairs of source and correct sentences, we extracted from each sentences all the nonidentical pairs (*помоему/по-моему, кто то/кто-то* and *то же/тоже* in the example above) and use these tokens for performance evaluation. We executed the same procedure on the pairs of source and

candidate sentences, where the candidate sentences are obtained from the correction sentences. We obtain two sets S_{corr} and S_{part} containing pairs of the form ((sentence number, source token), correction token) for source-correct and source-participant alignments. Then we calculated the number TP of true positives which is $||S_{corr} \cap S_{part}||$—the number of typo tokens properly corrected by the system. To obtain the precision score we divided this quantity by $|S_{part}|$ total number of corrections made by the system. The recall score was calculated as TP / $|S_{part}|$—the fraction of typo tokens, which were corrected properly. Note that false negatives in this case are both typos for which a wrong correction was selected and the typos left without correction.

We calculated F1-measure as the harmonic mean between precision and recall. All the three metrics were reported by the evaluation script; however, only F1-measure was used to rank the participants. In the final version of the evaluation script we also reported the percentage of correct sentences just for comparison.

When testing the alignment procedure, we found one subtle case not captured by the alignment algorithm. Consider the source sentence *«я не сколька не ожидал его увидеть»* (6) and its correction *«я нисколько не ожидал его увидеть»* (7). Suppose the spellchecker corrected both the mistyped words but did not manage to remove the space, yielding the sentence *«Я ни сколько не ожидал его увидеть»* (8). Literal application of the procedure above gives us two nontrivial alignment groups in the source-candidate pair: *«не/ни»* and *«сколька/сколько»*. Both these pairs were not observed in the reference alignment, therefore we obtain two false positives. Note that leaving the mistyped word *«сколька»* untouched yields better score since in this case only one unobserved aligned pair *«не/ни»* appears.

To improve this deficiency we made the following minor correction: we forced the alignment between source and suggestion sentences to have the same source components as in the source-correct alignment. For example, in the sentence above, the groups *«не/ни»* and *«сколька/сколько»* were joined together to obtain the pair *«не сколька/ни сколько»*, contributing one false positive instead of two.

3.3. Competition and participants

Seven research groups from 4 universities (MSU, MIPT, HSE, ISP RAS), 3 IT-companies (InfoQubes, NLP@Cloud, Orfogrammatika) and 2 cities (Moscow, Novosibirsk) successfully participated in the competition. These groups are listed in Table 1. Only best results from each group were taken into consideration and the number of attempts was not limited.

Table 1. Participants of SpellRuEval competition

Code of the group	scientific group
A	MIPT
B	GICR, MSU
C	HSE CompLing Spell
D	InfoQubes
E	ISP RAS
F	NLP@CLOUD
G	Orfogrammatika

4. Results and Discussion

All the systems presented in Table 2 used different toolkits and methods of automatic spelling correction, some of them are first time applied for Russian.

Table 2. Results of SpellRuEval competition

place	scientific group	Precision	Recall	F-measure	Accuracy
1	B	81.98	69.25	75.07	70.32
2	G	67.54	62.31	64.82	61.35
3	A	71.99	52.31	60.59	58.42
4	E	60.77	50.75	55.31	55.93
	BASELINE	55.91	46.41	50.72	48.06
5	C	74.87	27.99	40.75	50.45
6	D	23.50	30.00	26.36	24.95
7	F	17.50	9.65	12.44	33.96

We also evaluated a baseline system. Like several participant systems, it uses edit distance for search and combination of edit distance and n-gram model probability for ranking. In takes all the candidates on the distance of at most one edit from the source word and rank the obtained sentences using the sum of logarithmic edit distance score and language model score. Trigram language model was obtained using KenLM toolkit [Heafield et al., 2013] trained on the same data that was used by team B.

4.1. Methods and their efficiency

We collected the information about the methods and tools used by the competitors in Table 3 in the Appendix. Competition results show that all the teams used large dictionaries with approximately the same size, however, the difference in results is substantial. It means that the algorithms used for correction are more significant than additional data. It also proves that it is more important (and more difficult) to select a correct candidate than to find it, though several types of errors cannot be captured by basic search model based on edit distance without using additional errors lists or phonetic similarity. All the three top-ranked teams used a combination of edit distance and trigram language model for candidate ranking. It is interesting that morphological and semantic information gives no or little advantage in comparison with language model. One of the teams used Word2Vec instead of traditional n-gram model, which results in rather high precision (but not the best among all participants), though the recall was moderate in comparison with other results. It shows that Word2Vec is very successful in capturing frequent patterns, however, this method alone cannot detect all the errors. As expected, real-word errors were the most difficult to capture even by the top competitors, another source of difficult errors were misplaced hyphens and spaces. Probably, to correct such errors at least some rules of Russian orthography and grammar should be handcrafted, since such errors are too

frequent and subtle to be captured by pure statistical methods. Last but not the least, it is interesting that the competition winner (and actually the three best teams) used a rather simple algorithm: find the candidates using Levenshtein distance and rerank them using language model. It offers much room for future improvement by careful integration morphological, grammar or semantic features; however, it is not an easy task, as direct incorporation of morphology gave no advantage in current competition.

5. Conclusions

SpellRuEval 2016 has brought together a number of IT companies and academic groups that work on Russian Web text processing and normalization, so that it became possible to compare state-of-the-art methods in the field for Russian. The results have shown that the problem of automatic spelling correction for Russian social media is far from its solution. Up to now the best results are obtained using simple combination of edit distance candidate search and trigram language model, so future improvement can be achieved by adding morphological and semantic component to this basic framework.

The competition has the following practical outcomes:

- we have measured the current baseline of automatic spelling correction for Russian: on social media the baseline method show F1-measure of 50% and sentence accuracy of 48%. State-of-the-art methods used by the competition winner achieve F1-Measure of 75% and sentence accuracy of 70%.
- Various approaches to automatic spelling correction for Russian were tested; we have compared the role of different language models (ngram vs Word2Vec), different candidate search algorithms (dictionary lookup vs dictionary-free) and relative significance of different model elements (dictionary size, edit distance, language model, morphology and semantics usage). The results show that dictionary size is not the main factor, much more important is the adequacy of ranking model. Using more fine-grained features than simple edit distance score also improves the performance slightly. However, current system gain little or no advantage from morphological or semantic information which leaves much room for future improvement.
- the manually tagged golden standard set was developed, consisting of nearly 2000 sentences with different types of mistakes (typos, grammatical, orthographic, cognitive errors etc.) and their corrected variants. The organizers hope that the training set and golden standard (available at URL http://www.webcorpora.ru/wp-content/uploads/2016/01/source_sents.txt and http://www.webcorpora.ru/wp-content/uploads/2016/01/corrected_sents.txt) will help other researchers to evaluate their algorithms;

The experience of first SpellRuEval challenge could be useful for organizers and participants in future spell checking competitions. It would be interesting to test how linguistic information such as morphology, syntax or semantics could help in this task. The methods proposed could be also helpful in similar task like automatic grammar correction or social media text normalization. We hope to present one of these tasks in future Dialogue Evaluation competitions.

Acknowledgements

We would like to thank all colleagues who participated in the annotation of the golden standard. We also thank all the teams who took part in the competition, particularly to HSE and Orfogrammatika teams for their fruitful suggestions. We also would like to thank Eugeny Indenbom and Vladimir Selegey for their deep insight and helpful advice during the organization of the competition.

References

1. *Ahmad F., Kondrak G.* (2005) Learning a spelling error model from search query logs //Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing.—Association for Computational Linguistics,—pp. 955–962.
2. *Andrew Carlson and Ian Fette*. Memory-based context-sensitive spelling correction at web scale //Machine learning and applications. ICMLA 2007, Sixth international conference on.—IEEE, 2007.—pp. 166–171.
3. *Baytin A.* (2008), Search query correction in Yandex [Ispravlenie poiskovykh zaprosov v Yandekse], Russian Internet technologies [Rossijskie Internet-tekhnologii], 2008.
4. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.,* (2013), Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. Proceedings of Web as Corpus Workshop (WAC-8), Lancaster.
5. *Brill E., Moore R. C.* (2000) An improved error model for noisy channel spelling correction. In ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, pp. 286–293. Association for Computational Linguistics.
6. *Cucerzan S., Brill E.* (2004) Spelling correction as an iterative process that exploits the collective knowledge of web users. In Proceedings of EMNLP 2004, pp. 293–300.
7. *Dale R., Anisimoff I., Narroway G.* (2012) A Report on the Preposition and Determiner Error Correction Shared Task. In Proceedings of the NAACL Workshop on Innovative Use of NLP for Building Educational Applications.
8. *Dale R., Kilgarriff A.* (2011) Helping Our Own: The HOO 2011 Pilot Shared Task. In Proceedings of the 13th European Workshop on Natural Language Generation.
9. *Damerau F. J.* (1964) A technique for computer detection and correction of spelling errors. Communications of the ACM-7, pp. 171–176.
10. *Duan, H., Hsu, B.-J.* (2011) Online Spelling Correction for Query Completion. In Proceedings of the International World Wide Web Conference, WWW 2011, March 28–April 1, Hyderabad, India.
11. *Flor M.* (2012) Four types of context for automatic spelling correction //TAL.—T. 53.—№ 3.—pp. 61–99.
12. *Golding A. R., Roth D.* (1999) A winnow-based approach to context-sensitive spelling correction //Machine learning.—Vol. 34.—№ 1–3.—p. 107–130.
13. *Golding A. R., Schabes Y.* (1996) Combining trigram-based and feature-based methods for context-sensitive spelling correction //Proceedings of the 34th

annual meeting on Association for Computational Linguistics.—Association for Computational Linguistics,—pp. 71–78.
14. *Heafield K., Pouzyrevsky I., Clark J., Koehn I.* (2013) Scalable Modified Kneser-Ney Language Model Estimation //ACL (2).—pp. 690–696.
15. *Hirst G., Budanitsky A.* (2005) Correcting real-word spelling errors by restoring lexical cohesion //Natural Language Engineering.—Vol. 11.—№ 01.—pp. 87–111.
16. *Kernighan M. D., Church K. W., and Gale W. A.* (1990) A spelling correction program based on a noisy channel model. In Proceedings of the 13th conference on Computational linguistics, pages 205–210. Association for Computational Linguistics.
17. *Kukich K.* (1992) Techniques for automatically correcting words in texts. ACM Computing Surveys 24, pp. 377–439.
18. *Li M., Zhang Y. , Zhu M., Zhou M.* (2006) Exploring distributional similarity based models for query spelling correction //Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics.—Association for Computational Linguistics, 2006.—pp. 1025–1032.
19. *Liu V., Curran J. R.* (2007) Web Text Corpus for Natural Language Processing // EACL.—2006.
20. *Manning C.* (2011). Part-of-speech tagging from 97% to 100%: is it time for some linguistics? Proc. of CICLing.
21. *Mays E., Damerau F. J., Mercer R. L.* (1991) Context based spelling correction // Information Processing & Management. –Vol. 27.—№ 5.—pp. 517–522.
22. *McIlroy M. D.* (1982) Development of a Spelling List. AT&T Bell Laboratories.
23. *Mitton R.* (1987) Spelling checkers, spelling correctors and the misspellings of poor spellers //Information processing & management.—1987.—Vol. 23.—№ 5.—pp. 495–505.
24. *Mohit B., Rozovskaya A., Habash N., Zaghouani W., Obeid O.* (2014) The First QALB Shared Task on Automatic Text Correction for Arabic. In Proceedings of EMNLP Workshop on Arabic Natural Language Processing, Doha, Qatar, October.
25. *Ng H. T., Wu S. M., Briscoe T., Hadiwinoto C., Susanto R. H., Bryant C.* (2014) The CoNLL-2014 Shared Task on Grammatical Error Correction. In Proceedings of CoNLL: Shared Task.
26. *Ng H. T., Wu S. M., Wu Y., Hadiwinoto Ch., Tetreault J.* (2013) The CoNLL-2013 Shared Task on Grammatical Error Correction. In Proceedings of CoNLL: Shared Task.
27. *Oflazer K.* (1996) Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction //Computational Linguistics.—Vol. 22.—№ 1.—pp. 73–89.
28. *Panina M. F., Baitin A. V., Galinskaya I. E.* (2013) Context-independent autocorrection of query spelling errors. [Avtomaticheskoe ispravlenie opechatok v poiskovykh zaprosakh bez ucheta konteksta], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2013" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2013"], Bekasovo, pp. 556–568.
29. *Pedler J., Mitton R* (2010). A large list of confusion sets for spellchecking assessed against a corpus of real-word errors //Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10).

30. *Popescu O., Phuoc An Vo N.* (2014) Fast and Accurate Misspelling Correction in Large Corpora. Proceedings of EMNLP 2014: Conference on Empirical Methods in Natural Language Processing. Doha, Qatar.
31. *Ristad E. S., Yianilos P. N.* (1998) Learning string-edit distance //Pattern Analysis and Machine Intelligence, IEEE Transactions on.—Vol. 20.—№ 5.—p. 522–532.
32. *Rozovskaya A., Bouamor H., Habash N., Zaghouani W., Obeid O., Mohit B.* (2015) The Second QALB Shared Task on Automatic Text Correction for Arabic. ANLP Workshop 2015, The Second Workshop on Arabic Natural Language Processing, July 30, 2015 Beijing, China.
33. *Rozovskaya A., Roth D.* (2013) Joint learning and inference for grammatical error correction //Urbana.—Vol. 51.—pp. 61801.
34. *Schaback J., Li F.* (2007) Multi-level feature extraction for spelling correction // IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data.—pp. 79–86.
35. *Schulz K., Mihov S.* (2002) Fast string correction with Levenshtein automata // International Journal on Document Analysis and Recognition.—Vol. 5.—№ 1.—pp. 67–85.
36. *Shavrina T., Sorokin A.* (2015) Modeling Advanced Lemmatization for Russian Language Using TnT-Russian Morphological Parser. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2015", RSUH, Moscow.
37. *Toutanova K., Moore R. C.* (2002) Pronunciation modeling for improved spelling correction //Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.—Association for Computational Linguistics—pp. 144–151.
38. *Vladimir I. Levenshtein,* (1965) Binary codes capable of correcting deletions, insertions, and reversals [Dvoichnye kody s ispravleniem vypadenij, vstavok i zameschchenij simvolov], Doklady Akademij Nauk SSSR.—1965.—Vol. 163.—pp. 845–848.
39. *Wang K., Pedersen J.* (2011) Review of MSR-Bing web scale speller challenge. In Proceeding of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 1339–1340.
40. *Wang K., Thrasher C., Viegas E, Li X., Hsu B. H.* (2010) «An overview of Microsoft Web N-gram corpus and applications.» In Proceedings of the NAACL HLT 2010 Demonstration Session, pp. 45–48.
41. *Whitelaw C., Hutchinson B., Chung G. Y, Ellis G.* (2009) Using the web for language independent spellchecking and autocorrection. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2, pp. 890–899. Association for Computational Linguistics.
42. *Corpus of Russian Student Texts* web-corpora.net/CoRST/search/index.php?interface_language=ru.
43. *Hunspell:* open source spell checking, stemming, morphological analysis and generation, 2011. http://hunspell.sourceforge.net/.
44. *Russian National Corpora* http://www.ruscorpora.ru/search-main.html.
45. *Russian OpenCorpora* http://opencorpora.org/.
46. *Spelling Alteration for Web Search Workshop.* 2011. Bellevue, USA http://spellerchallenge.com/Workshop.aspx.

Appendix 1. Analysis of methods used by SpellRuEval participants

Group code	Methods used	Dictionary size: wordforms	Dictionary size: lemmata	Error detection	Error correction	Most common mistakes of the system:	Training set (excluding SpellRuEval training set)
A	Edit distance for candidate search and language model for selection	3,700,000	230,000	Dictionary-based by edit distance, keyboard adjacency and graphic similarity	Correction is selected by trigram language model	wrong candidate ranking, syntactic errors, real-word errors	Corpus of 213,000,000 words with morphology for language model
B	Multi-level spelling correction, error model, n-grams, POS-tag n-grams	3,700,000	230,000	Each word is a potential error, the best variant is selected by the correction score	Edit distance and phonetic coding for candidate search, trigram language model and error model for ranking	semantic errors, wrong candidate ranking	50 million tokens from Live Journal
C	Word2Vec, n-grams, hybrid error model combining: 1) traditional channel model that uses single letter edits, 2) the model introduced by Brill and Moore, 3) extended version of the channel model with wider context edits	no dictionary used	no dictionary used	error detection confidence classifier, uses word2vec filtered vector scores	Autocorrector that processes words flagged as misspellings by the classifier	wrong candidate ranking, syntactic errors, undetected errors, real-word errors	Corpus of Russian Student Texts (CoRST) 2.5 million tokens + 13.8 million tokens from Blogs + 22.2 million tokens from newspapers for n-gram model
D	edit distance, automatic rule-based paradigm construction	5,000,000	260,000	Dictionary based + dictionary look-up with suitable suffixation rules	Edit distance. If there are 2 or more candidates with same distance, choice is random	wrong inflection, wrong candidate ranking	No extra-data
E	n-grams, rule-based dictionary look-up	5,095,000	390,000	Dictionary based (OpenCorpora Dictionary)	Candidate ranking with 3-grams from training set and edit distance	wrong candidate ranking, spacing errors	400 million tokens from newspapers, social networks and WikiPedia
F	chunking in dependency model, vector model, Jflex, Apache Tika	5,095,000	390,000	Dictionary based (OpenCorpora Dictionary)	Ranking the ChunkTrees with max.number of words in the sentence, then correcting wordform depending on syntactic and POS-tags.	undetected errors, orthographic errors	No extra-data used
G	n-grams, POS-tag n-grams	5,500,000	400,000	Dictionary-based	Candidate ranking with wordform 3-grams, POS-tag 2- and 3-grams, wordform frequency etc.	wrong candidate ranking, word separation errors	RNC sample of 1 million tokens with resolved homonymy

AUTOMATIC DETECTION OF MORPHOLOGICAL PARADIGMS USING CORPORA INFORMATION

Sorokin A. A. (alexey.sorokin@list.ru)[1,2],
Khomchenkova I. A. (irina.khomchenkova@yandex.ru)[1]

[1]Lomonosov Moscow State University, Moscow, Russia

[2]Moscow Institute of Physics and Technology, Dolgopudnyj, Russia

This paper deals with automatic induction and prediction of morphological paradigms for Russian. We apply a method of longest common subsequence to extract abstract paradigms from inflectional tables. Then we experiment with the automatic detection of paradigms using a linear classifier with lexeme suffixes and prefixes as features. We show that Russian noun paradigms could be automatically detected with 77% accuracy per paradigm and 93% accuracy per word form, for Russian verbs per-paradigm accuracy reaches 76% and per-form accuracy is 89%. Usage of corpora information and character n-grams allows to improve these results up to 82% and 95% for nouns and 86% and 95% for verbs.

Keywords: abstract paradigm, paradigm induction, longest common subsequence, automatic paradigm detection, corpora-based paradigm detection

АВТОМАТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ МОРФОЛОГИЧЕСКИХ ПАРАДИГМ С ИСПОЛЬЗОВАНИЕМ КОРПУСНОЙ ИНФОРМАЦИИ

Сорокин А. А. (alexey.sorokin@list.ru)[1,2],
Хомченкова И. А. (irina.khomchenkova@yandex.ru)[1]

[1]Московский государственный университет им. М. В. Ломоносова, Москва, Россия

[2]Московский физико-технический институт, Долгопрудный, Россия

Данная работа посвящена автоматическому определению и классификации морфологических парадигм для русского языка. Абстрактные морфологические парадигмы выделяются с помощью метода наибольшей общей подпоследовательности. Основная часть работа посвящена проблеме вычисления полной парадигмы для неизвестной лексемы, для чего применяется линейная классификация. В качестве признаков для классификации используются префиксы и суффиксы данной лексемы. Мы показываем, что абстрактная парадигма может быть определена с точностью 77% для существительных и 76% для глаголов, в то время как точность по словоформам достигает 93 и 89%. В работе вводится новый алгоритм автоматического определения морфологической парадигмы, использующий корпусную информацию. Он позволяет достичь качества в 82% для именных и 86% для глагольных парадигм, в то время как точность по словоформам в обоих случаях становится равной 95%.

Ключевые слова: морфологическая парадигма, абстрактная парадигма, автоматическое определение парадигм, автоматическая классификация парадигм, корпусной метод определения парадигм

1. Introduction

The automatic induction and learning of morphological paradigms is very popular in the last years. State-of-the-art works include [Ahlberg et al., 2015] and [Nicolai et al., 2015], but several other papers are also worth mentioning (Ahlberg et al., 2014], [Durrett, DeNero, 2013]). This task has various applications, e.g. synthesis of surface word forms in machine translation and the automatic extension of morphological resources, such as wiktionary.org. The methods developed for paradigm learning can also be used in the automatic morphological analysis, e.g. for POS-tagging or lemmatization.

The automatic induction of morphological paradigms has a long history in the Russian linguistic tradition. The seminal work of A. A. Zaliznyak "Russkoe imennoe slovoizmenenie" [Zaliznyak, 2002] solves exactly this problem: how the complete description of morphological inflection could be recovered from empirical data. If we reconsider the algorithm of Zaliznyak from the computational point of view and omit the technical details specific to Russian phonology, it is essentially based on the method of longest common subsequence (LCS): the invariant part of inflected forms of the same lexeme is exactly their LCS. The method of LCS for automatic induction of morphological paradigms was reintroduced in works of Ahlberg, Hulden et al. ([Ahlberg et al., 2014], [Ahlberg et al., 2015]). However, for the purposes of computational linguistics, automatic induction of morphological paradigms from inflected tables is only the preliminary step. A more important question is how to detect the paradigm label and hence the complete inflectional table using only the base form of the lexeme. This problem is solved by machine learning techniques, using substrings of the source lexeme (e.g., its prefixes or suffixes) as features for the classifier.

There are practically no works on automatic detection of morphological paradigms for Russian: [Ahlberg et al., 2015] contains some results for noun declension but the quality of the source data is too low to consider them significant. We reimplement

the method of Hulden for paradigm induction with several technical modifications and use a linear classifier to derive these paradigms automatically from the lexeme. Our algorithm is able to recover complete morphological paradigm both for Russian nouns and verbs with accuracy of 77% for paradigms and 93 and 88% for word forms respectively. We also demonstrate that the usage of corpora information improves the percentage of correctly predicted paradigms up to 82% for nouns and 86% for verbs.

2. Abstract paradigms

For the compressed representation of morphological inflection we use the notion of an abstract paradigm, introduced in [Ahlberg et al., 2014]. From the mathematical point of view, a paradigm is a tuple of functions $F = \langle f_1, \ldots, f_n \rangle$ taking the same variables $x_1, \ldots, x_r \in \Sigma^+$, where $f_i(x_1, \ldots, x_r)$ operates from $(\Sigma^+)^r$ to Σ^+ ([Ahlberg et al., 2014], see also [Zaliznyak, 2002]). Here Σ is the finite alphabet and Σ^+ denotes the set of all words over this alphabet. Each of the functions f_i corresponds to some grammatical meaning c_i, the functions in set F are arranged according to a fixed order $c_1, \ldots, c_n$ of possible grammatical meanings. Literally speaking, a paradigm is a mapping from variables to strings. We use the term "abstract paradigm" to represent morphological paradigms formally. An abstract paradigm is a tuple of strings containing variables $x_1, x_2, \ldots, x_n$ (the variables are the same for all strings and have the same order elsewhere) and constant fragments, which are the same for all lexemes satisfying the given paradigm. These constant fragments vary between the forms of the same lexeme. On the contrary, the variables have the same value for all inflected forms but differ from lexeme to lexeme.

Let us explain these formal terms on a short example. Consider the declension tables of two Russian nouns *кусок* and *песок*. The paradigm function F is the same for both of them; in the first case it takes the variables $x_1 =$ кус and $x_2 =$ к, in the second one—$x_1 =$ пес, $x_2 =$ к.

Tab. 1. Abstract paradigm: an example

Grammeme	Pattern	F (кус, к)	F (пес, к)
Nom.Sg	x_1+о+x_2	кусок	песок
Nom.Pl	x_1+x_2+и	куски	пески
Gen.Sg	x_1+x_2+а	куска	песка
Gen.Pl	x_1+x_2+ов	кусков	песков
Dat.Sg	x_1+x_2+у	куску	песку
Dat.Pl	x_1+x_2+ам	кускам	пескам
Acc.Sg	x_1+о+x_2	кусок	песок
Acc.Pl	x_1+x_2+и	куски	пески
Instr.Sg	x_1+x_2+ом	куском	песком
Instr.Pl	x_1+x_2+ами	кусками	песками
Pr.Sg	x_1+x_2+е	куске	песке
Pr.Pl	x_1+x_2+ах	кусках	песках

Given the variable values, an abstract paradigm unambiguously determines the complete inflectional table. When a pattern and a word form are known, usually there is only one way to fit the pattern to the word: for example, the word *мешок* and the pattern x_1+о+x_2 yield a single combination of variable values x_1=меш, x_2=к. Nevertheless, applying the same pattern to the word *носок* results in two variants x_1=н, x_2=сок and x_1=нос, x_2=к. If we take into account several possible patterns, the number of decompositions can grow up dramatically. However, the variables are extracted not from a single word form, but from all the paradigm elements simultaneously, which restricts the set of possible combinations.

3. Longest common subsequence

Consider again the abstract representation of morphological paradigms. If we substitute strings of letters for the variables, these strings form a common subsequence of all generated words. In order to capture as much common material as possible, that subsequence should be the longest one. Therefore, the problem of paradigm detection has been reduced to the task of finding the longest common subsequence. We are not going to discuss the linguistic relevance of this approach and use it only as an empirical procedure. However, several important questions emerge:

1. How to calculate the longest common subsequence algorithmically?
2. What subsequence to select when several subsequences have the same length?
3. How to extract variable values when the LCS is known?

For the first task we use finite automata. It is straightforward to construct an automaton recognizing all the common subsequences of given strings and then extract the longest word this automaton accepts (we omit algorithmical details). Although, this automaton could be nondeterministic and an equivalent deterministic state automaton may have much larger number of states (up to 2^n where n is the number of states of initial nondeterministic automaton). To prevent this exponential growth we bound the length of gaps between the consequent letters of the subsequence, as well as the gap before the first letter of the subsequence. This limitation is also justified from the linguistic point of view: consider two verb forms *разместиться* and *размещусь*, their LCS *размес* has length 6. However, *с* in the LCS is an artifact of the method, not an element of common stem. Besides, alterations like *ст/щ* are among the phenomena which are difficult to capture by LCS algorithm.

The construction of finite automata recognizing all common subsequences for the words *мотюк* and *окот* is illustrated below. The edges contain not only the symbols, but also the positions of these symbols in the words. This trick allows to simplify the extraction of an abstract paradigm from the LCS.

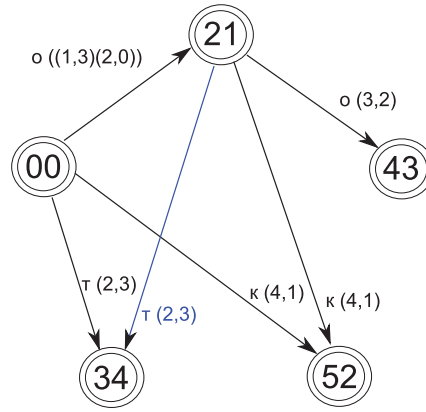


Fig. 1. DSA for common subsequences of the word *мотоκ* and *окот*

In the example above there are 3 longest common subsequences: *оо*, *ок*, *от*. Possible variants of their positioning are shown in the table below.

Tab. 2. LCS for the words *мотоκ* and *окот*

LCS	LCS positioning variants
о-о	м**от**о**к**, **о**к**от**
о-т	м**от**о**к**, **о**к**от**
о-т	м**от**о**к**, о**от**
о-к	м**от**о**к**, **о**к**от**
о-к	м**от**о**к**, **о**к**от**

Already in this artificial example there are multiple variants for LCS positioning. The same problem emerges in practice: consider a partial declension table of the word *песок*. There are two candidates for the LCS: *пес-о* and *пес-к* both of length 4.

Tab. 3. Ambiguous LCS positioning: an example

Nom.Sg	**песок**	Nom.Sg	**песок**
Gen.Pl	**песк**ом	Gen.Pl	**песк**ом
Instr.Sg	**песк**ов	Instr.Sg	**песк**ов

We use two heuristics for disambiguation: the first selects the variant with the minimal number of variables (variables are the maximal contiguous parts of the LCS). However, this heuristic does not give us a solution here: both subsequences consist of two variables. Then we apply our second heuristic: choose the variant with the least total length of gaps. Then the variant **песок**-**песк**ов-**песк**ом is preferred, since it leads to a single gap of length 1 while its counterpart generates two such gaps (of total length 2).

4. Automatic detection of paradigms

In the previous section we have discussed the algorithm for morphological paradigms induction. However, it is not a central problem of the paper; we are mainly interested in the automatic detection of such paradigms for unknown words. We consider the following task: given an unknown word of a known part-of-speech (say, a noun *арка*), determine its complete declension table. The algorithm selects one of many potential variants, several of which are listed in Table 4.

Tab. 4. Multiple possible paradigms for the word *арка*

Paradigm	Variables
1#1+ы#1+а#1+ов#1+у#1+ам#1#1+ы#1+ом#1+ами#1+е#1+ах	1=арка
1+а#1+а#1+ы#1#1+е#1+ам#1+у#1+ы#1+ой#1+ами#1+е#1+ах	1=арк
1#1+ы#1+а#1+ов#1+у#1+ам#1#1+а#1+ов#1+ами#1+е#1+ах	1=арка
1+2+а#1+2+и#1+2+и#1+о+2#1+2+е#1+2+ам#1+2+у#1+2+и #1+2+ой#1+2+ами#1+2+е#1+2+ах	1=ар, 2=к

We may attempt to recover a correct paradigm using deterministic rules such as "when a noun ends with *а* then this *а* is a flection, not a part of a stem" (counterexample: *баккара*) and if such word ends with "*Cка*" for some consonant *C* then *о* is inserted between *C* and *к* in genitive plural (counterexample: *ласка*). However, all such rules have counterexamples and their manual design is a very labour-intensive task. Therefore we have decided to learn inflectional patterns automatically applying algorithms of machine learning. We use as features all the suffixes[1] whose length does not exceed the given maximum (say, 5). The suffixes are encoded as binary indicators; for example, the word *учитель* is described by a binary vector with five nonzero elements, corresponding to suffixes *-ь*, *-ль*, *-ель*} etc. (see Table 5 below). The absence of a suffix in the training set is encoded by a special placeholder, in this case longer suffixes are not taken into account since they were not observed in the training set either. For example, if the suffix *-ль* was preceded only by *е* in the training set, then both words *мораль* and *фасоль* are encoded by a vector containing three ones for suffixes *-ь*, *-ль* and *!ль*} where *!* denotes an unobserved letter.

Tab. 5. Feature encoding scheme

	а	к	ка	ла	ик	рка	…
арка	1	0	1	0	0	1	…
школа	1	0	0	1	0	0	…
блик	0	1	0	0	1	0	…
…	…	…	…	…	…	…	…

[1] We use the term "suffix" ("prefix") for an arbitrary substring in the end (in the beginning) without any regard to morphology

Since prefixes carry no information about noun morphology, we do not use them as features for noun paradigm prediction. In the case of verbs, conversely, they can be used to determine verb aspect. If d is the maximal length of suffixes used as features, then the number of possible features grows roughly exponentially with d and may reach 20,000 for $d = 5$. To reduce training time and remove noisy features we retain only a fixed percentage of the most unambiguous features. As the measure of ambiguity for the feature f_j we take $\max_i P\big(c(L) = c_i \big| f_j(L) = 1\big)$—the probability of the most frequent class provided f_j is present. We also remove features which appear less than 3 times in the training set.

5. Evaluation of paradigm classifier

We have evaluated our approach on Russian verbs and nouns. For both tasks we took 5,000 most frequent words of the corresponding part of speech from the dictionary of Lyashevskaya and Sharoff ([Lyashevskaya, Sharoff, 2009]). We automatically downloaded complete inflectional tables from the Wiktionary (ru.wiktionary.org). For nouns the tables contained at most 12 items for 6 cases and 2 numbers (several cells in the paradigm could be empty, e.g. for pluralia tantum). Sometimes the cell contained two values (for example, Instr.Sg. of first declension nouns), in this case we always chose only the first form. We extracted 239 abstract paradigms for noun declension, 69 of them contain more than 5 examples and 108—only a single example. 10 most frequent paradigms are listed in Table 11 of the Appendix.

In the case of verbs typical Wiktionary form for imperfect aspect contains 21 simple forms (https://ru.wiktionary.org/wiki/\%D0\%B6\%D0\%B5\%D0\%BB\%D0\%B0\%D1\%82\%D1\%8C) including infinitive and omitting composite future form and empty cells. For paradigm induction we used only 13 of them: 6 present forms, 4 past, 2 forms of the imperative and the basic infinitive form. Even in such restricted form verb conjugation demonstrate more irregularities then noun declension, so the sample of 5,000 verbs contains 305 paradigms with 120 of them having 5 or more representatives and 92—a single representative. 10 most frequent paradigms are shown in Table 12. We bound maximal gap length by 2, therefore the algorithm does not recognize *с* as part of the LCS in the examples like *играться/играешься/играйтесь*.

In our experiments we randomly separated the sample on 2 equal halves, using one for testing and the other one for training. The results were averaged for 5 random splits. In the case of nouns we did not use prefixes as features and bound suffix length d by 3, 5 or 7. The percentage p of selected features was 0.10, 0.25 or 0.5. In the case of verbs we calculated the suffix length without the reflexive affixes *-ся* and *-сь*. We also used the prefix features with the maximal length of 2 for verb conjugation. To predict paradigm labels we used the logistic regression classifier from sklearn package [Pedregosa et al., 2011], which itself uses the LIBLINEAR library [Fan et al., 2008]. The results are presented in Table 6 and Table 7. We report both per-paradigm (the percentage of correctly predicted abstract paradigms) and per-form (the fraction of correct word forms) accuracy.

Tab. 6. Prediction accuracy for noun paradigms classification

	0.1		**0.25**		**0.5**	
3	77.19	93.47	77.26	93.47	77.25	93.47
5	77.38	93.50	77.32	93.48	77.32	93.48
7	77.44	93.45	77.35	93.43	77.35	93.43

Since the result of nouns is practically independent from the classifier parameters, we fix $p = 0.1$ and $d = 5$ in future experiments. We use the same setting for the verbs task, however, in this case the impact of feature length is more significant.

Tab. 7. Prediction accuracy for verb paradigms classification

	0.1		**0.25**		**0.5**	
3	51.41	79.96	51.41	79.96	51.41	79.94
5	76.30	88.83	76.09	88.62	75.94	88.62
7	77.06	88.36	78.01	89.35	77.96	89.38

We also study how the prediction quality changes with the size of the training set. When there is little training data available, a lemma may not fit to all inflectional patterns observed in training phase (say, a verb ends with *-ти* and all the infinitives in the training set ended with *-ть*, *-ться* or *-чь*). In such cases we allow the system consult a complete list of paradigms, no matter whether they were observed in training. The dependence between training data size and system performance is shown in Table 8.

Tab. 8. Train data percentage and performance quality

	Training data fraction					
Task	**0.1**	**0.25**	**0.5**	**0.6**	**0.7**	**0.8**
Nouns	71.76 91.15	75.05 92.32	77.38 93.50	77.95 93.70	77.88 93.77	77.40 93.84
Verbs	65.50 83.83	71.50 86.27	76.30 88.83	77.49 89.36	77.60 89.41	77.56 89.50

6. Analysis of results

It is uninformative to compare results for different languages and even for different datasets. As we know, the only experiment on paradigm detection for Russian nouns was conducted by Ahlberg et al. in [Ahlberg et al., 2015], showing per-table accuracy of 66% and per-form accuracy of 89%. However, they used data collected from Freeling ([Padro, Stanilovsky, 2012]), which is of much lower quality than ours. They also used 5-fold cross-validation for performance evaluation, which means that

80% was left for training instead of only 50% in our experiment. However, the results for other languages, such as Catalan, French or Italian, reported in [Ahlberg et al., 2015] are much higher with per-table accuracy of over 90%. We claim that corpus-free methods are incapable of reaching comparable accuracy on Russian data due to the objective linguistic factors.

There are two main sources of errors in the case of noun paradigm prediction: the first is animacy/inanimacy affecting the forms of accusative, the second is *-а/-ы* in the form of Nom.Pl. In both cases the correct category does not depend on the surface form (consider *волчонок* vs *бочонок* or *голос* vs *колос*). The system also fails to discriminate between masculine and feminine nouns ending with *ь* (*мозоль* vs *король*). It is obvious that these ambiguities cannot be resolved without corpus statistics. We discuss this question in details in the next section.

For verbs the problem is more subtle. Often the mistake happens for the forms of imperative mood, for example, **тревожи* is predicted instead of *тревожь* or **похити* for *похить*. In such cases the forms or indicative mood are usually correct. Another common source of mistakes are *е/ё* in verb flections (compare *хлопнуть* and *толкнуть*). In this case the flection depends on the stress position in the infinitive form, however, we removed the stress signs in our data since they are marked inconsistently in Wiktionary itself. Such mistakes affect only several forms (imperative or third person present tense). Errors of the second type touch practically all forms of the paradigm. It often happens for the verbs ending on *-ать* (*венчать* vs *кричать*). The system also fails in the case of phonetic alterations (*унизить/унижу*), especially when they happen inside the stem (*звать/зову* or *слать/шлю*).

Summarizing, the spectrum of possible errors for Russian verb paradigm prediction is wider than for Russian nouns, which explains lower per-form quality in the verb prediction task. However, in both cases more training data does not help, as shown in Table 8. We consider the sources of additional information in the next section.

7. Corpus-based methods of paradigm predictions

In this section we experiment with other features which might be helpful for automatic paradigm detection. In the verb paradigm task incorrectly predicted forms sometimes violate the rules of Russian phonology like in **осуществься* or **исчежь*} for *исчезни*. These incorrect forms might be rejected if we extend the model by phonological features. This idea is realized as following:

First, we train a character n-gram model on the training data. Then we augment the algorithm with second classifier on the top of the first. It takes as features logarithmic probabilities predicted by the classifier on the first level as well as the scores of the language model. If the basic classifier has predicted c_i as paradigm label for the lemma L, we generate all the forms $w_{i,1}, ..., w_{i,m}$ of this lexeme according to the paradigm; then we take as language model score the averaged sum

$$s(L, c_i) = \frac{\sum_{j=1}^{m} - \log P_{lm}(w_{i,j})}{m}$$

where $P_{lm}(w_{i,j})$ is the probability of word form w_i, according to character ngram model. We test two ways of accomodating the language model log-scores: in the first case we use them as features of the linear classifier. In the second variant we used language model scores only for filtering, discarding a paradigm c_i if its score $s(L, c_i)$ is greater than $s_0 + \alpha$ where s_0 is the lowest value among $s(L, c_i)$ and α is some redefined constant. We used 5-gram language models trained on the set of word forms from the training data and smoothed the model counts using Witten-Bell smoothing ([Chen, Goodman, 1996]). The results for Nouns and Verbs tasks are presented in Table 9, we used p=0.1 and d=5 for feature fraction and suffix length in all trials, the percentage of training data was again 0.5.

Tab. 9. Using character model for paradigm prediction

Task	No character scores	Character scores as features	Character scores as filters
Nouns	77.38 93.50	77.42 93.50	77.36 93.42
Verbs	76.30 88.86	80.37 90.92	77.01 89.35

We observe that language model has no effect for the Nouns task. On the contrary, on the verbs task filtering already improves performance significally, while combining language model scores with initial paradigm probabilities increases prediction quality by 3 percents more. It is easy to explain since the main source of errors for nouns was the confusion between animate/inanimate nouns where both the predictions are phonologically plausible. Conversely, in the Verbs task the mispredicted forms in imperative like **осуществься* has low probability according to character n-gram models which allows the system to exclude them.

The main contribution of our paper is corpora-based algorithm for paradigm prediction. Again, we accommodate corpora counts together with the logarithmic probabilities predicted by the basic classifier on the second stage of our algorithm. More precisely, after generating the word forms $w_1, ..., w_m$ of the lexeme L according to hypothetic paradigm c_j, we calculate the corpus score by the formula $C = \sum_{j=1}^{m} - \log C(w_j)$, where $C(w_j)$ is the number of times w_j occurs in the corpora. All counts are incremented by 1 to avoid zero probabilities. This method resembles the method of [Ahlberg et al., 2014], however, we make one modification to deal with homonymy: if a word form occurs two times in the paradigm (for example, in nominative and genitive), then we divide all the corpora counts of it by 2. Without this modification, this algorithm favours invariable nouns.

However, we are still unable to discriminate between inanimate and animate nouns by our algorithm since the set of word forms is the same in both cases. The only difference is that genitive forms of animate nouns would be more frequent than the ones of inanimate since they appear in accusative also. To capture this difference we should measure the similarity between the expected distribution of case forms and the observed proportion of their counts. Let $P = [p_1, ..., p_m]$ be the expected probabilities of different word forms according to their grammemes and $N = [N_1, ..., N_m]$ be their observed counts. We normalize the empirical distribution by its sum $N = \sum_j Nj$,

obtaining the empirical probability distribution $Q = [q_1, \ldots, q_m]$, where $q_j = \frac{N_j}{N}$. Then the difference score equals

$$D(\mathcal{N}, \mathcal{P}) = \sum_j q_j \log \frac{q_j}{p_j} \cdot \log N$$

Note that this measure is simply Kullback-Leibler divergence between Q and P multiplied by the log count of the given lexeme. The expected form counts were collected in the training phase separately for each paradigm. The results for corpora-based paradigm prediction are shown in Table 10. We used the counts from Russian National Corpora available on ruscorpora.ru/corpora-freq.html.

Tab. 10. Using character model for paradigm prediction

Task	No corpora	Corpora counts as features	Counts and divergences as features
Nouns	77.38 93.50	80.21 95.34	82.73 95.67
Verbs	76.30 88.83	84.30 93.81	83.66 93.73

We observe that using corpora counts indeed leads to a substantial gain in performance in both tasks. However, in the case of verbs most of the advantage is obtained from corpora counts themselves, using similarity scores slightly worsens performance. On the Nouns task similarity scores, on the contrary, leads to a further improvement in per-table accuracy. Indeed, the most difficult problem for nouns is animacy/inanimacy differentiation where absolute counts are useless. In the verb tasks, conversely, homonymy plays no role, therefore, similarity scores are redundant and make the data noisier.

Inspecting remaining incorrect predictions, we found that in the Verbs task they are mainly caused by wrong imperative form generation. Often corpus counts cannot resolve this problem because imperative forms are not very frequent for many verbs: both *кровоточи* and **кровоточь* do not appear in the RNC counts. Often corpora features are not powerful enough to overcome the gap caused by first level classifier. For example, for the verb *лгать* the correct paradigm has probability 0.01 after the first stage. Joint classifier raises it up to 0.3, however, it is too low to rank this hypothesis on the top. The same problem arises in the task of noun paradigm prediction: for most of the erroneous predictions the correct paradigm was excluded already by the basic classifier or obtained an extremely low probability.

We also combined character n-gram scores with the corpora-based classifier, which improved the performance further. For the Nouns task the gain was marginal (82.80% instead of 82.73% for per-table accuracy), however, the accuracy of paradigm prediction for verbs achieved 86.51% instead of 84.30%. The per-form accuracy also increased significantly, reaching 95.66% in comparison with 93.81%.

8. Conclusion

We have developed a system for automatic paradigm induction and prediction. Our algorithm of paradigm induction is based on the method of longest common subsequence. To predict paradigms automatically we apply a logistic regression classifier using suffix and prefix features. This classifier achieves accuracy of 77% on Russian nouns and 76% on Russian verbs in paradigm prediction task, the percentage of correctly predicted forms is 93% and 88% respectively. We have also designed a corpora-based algorithm of paradigm prediction using the basic classifier on its first stage. This improves the accuracy of paradigm prediction to 82% on nouns and 86% on verbs, per-form accuracy reaches 95 % for both tasks. These results are substantially better than previously achieved for Russian in [Ahlberg et al., 2015] (the authors of that work used another dataset and experiment setting).

We plan to improve our results further by using corpora information more extensively. Our results show that taking into account relative frequencies of grammemes enhances the quality of corpora-based methods. Therefore modelling the distribution of grammemes more accurately should leave to further improvement. For this goal we plan to use morphologically disambiguated corpora. Another improvement could be achieved by grouping together the corpus statistics for the words of presumably the same paradigm.

Our results could be used for automatic morphological analysis and synthesis in such tasks as POS-tagging or lemmatization. Modern techniques of lemmatization such as used in [Jonjejan, Dalianis, 2009] also use the LCS approach but apply it to each word form separately without using full inflectional table. Our method incorporates information from the whole paradigm, therefore it could potentially improve state-of-the-art algorithms of morphological analysis for Russian. Since our system does not predict the best inflectional table only, but returns the probabilities of possible paradigms, it can be used as a component of a joint classifier, taking into account context model probabilities as well as single word scores. Using context information together with suffix/prefix features could also help to determine word part-of-speech, which is a preliminary step for our algorithm.

This task is especially important for Web texts, which contain numerous out-of-vocabulary words whose inflection cannot be determined by dictionary-based methods. We plan to test our approach for morphological processing of social media texts in future studies.

References

1. [Ahlberg et al., 2014] *Ahlberg M., Forsberg M., Hulden M.* (2014) Semi-supervised learning of morphological paradigms and lexicons // EACL 2014, p. ~569.
2. [Ahlberg et al., 2015] *Ahlberg M., Forsberg M., Hulden M.* (2015) Paradigm classification in supervised learning of morphology // Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies (NAACL-HLT 2015), Denver, CO, pp. 1024–1029.
3. [Chen,Goodman, 1996] *Chen S. F., Goodman J.* (1996) An empirical study of smoothing techniques for language modeling // Proceedings of the 34th annual meeting on Association for Computational Linguistics, pp. ~310–318.

4. [Durrett, DeNero, 2013] *Durrett G. and DeNero J.* (2013) Supervised Learning of Complete Morphological Paradigms. // HLT-NAACL, pp. 1185–1195.
5. [Fan et al., 2008] *Rong-En Fan R.-E., Chang K.-W., Hsieh C.-J., Wang X.-R., Lin C.-J.* (2008) LIBLINEAR: A Library for Large Linear Classification // Journal of Machine Learning Research, Vol. ~9, pp. ~1871–1874.
6. [Jonjejan, Dalianis, 2009] *Jongejan B., Dalianis H.* (2009) Automatic training of lemmatization rules that handle morphological changes in pre-, in-and suffixes alike. // Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Vol. ~1, pp. ~145–153.
7. [Lyashevskaya,Sharoff, 2009] *Lyashevskaya O. and Sharoff S.* (2009) Frequency dictionary of modern Russian language [Chastotnyj slovar› sovremennogo russkogo yazyka], Azbukovnik, Moscow.
8. [Nicolai et al., 2015] *Nicolai G., Cherry C., Kondrak G.* (2015) Inflection Generation as Discriminative String Transduction // Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies (NAACL-HLT 2015), Denver, CO, pp. 923–931.
9. [Padro, Stanilovsky, 2012] *Padro L., Stanilovsky E.* (2012) FreeLing 3.0: Towards Wider Multilinguality // Proceedings of the Language Resources and Evaluation Conference (LREC 2012), Istanbul, Turkey, pp.~2473–2480.
10. [Pedregosa et al., 2011] *Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E.* (2011) Scikit-learn: Machine Learning in Python. // Journal of Machine Learning Research, Vol. ~12, pp. ~2825–2830.
11. [Zaliznyak, 2002] *Zaliznyak A. A.* (2002) Russian nominal inflection with a supplement of selected works on modern Russian and general linguistics. [Russkoe imennoe slovoizmenenie s prilozheniem izbrannyh rabot po sovremennomu russkomu yazyku] Yazyki slavianskoj kul'tury, 2002.

Appendix

Tab. 11. Most frequent abstract paradigms for Russian nouns

№	Abstract paradigm	Count	Example
1	1#1+ы#1+а#1+ов#1+у#1+ам #1#1+ы#1+ом#1+ами#1+е#1+ах	959	0=аборт, 1=аборт
2	1+е#1+я#1+я#1+й#1+ю#1+ям #1+е#1+я#1+ем#1+ями#1+и#1+ях	622	0=Евангелие, 1=Евангели
3	1+а#1+ы#1+ы#1#1+е#1+ам #1+у#1+ы#1+ой#1+ами#1+е#1+ах	444	0=автомашина, 1=автомашин
4	1+ь#1+и#1+и#1+ей#1+и#1+ям #1+ь#1+и#1+ью#1+ями#1+и#1+ях	330	0=активность, 1=активност

№	Abstract paradigm	Count	Example
5	1+я#1+и#1+и#1+й#1+и#1+ям #1+ю#1+и#1+ей#1+ями#1+и#1+ях	270	0=авария, 1=авари
6	1#1+ы#1+а#1+ов#1+у#1+ам #1+а#1+ов#1+ом#1+ами#1+е#1+ах	249	0=абонент, 1=абонент
7	1+2+а#1+2+и#1+2+и#1+о+2#1+2+е #1+2+ам#1+2+у#1+2+и#1+2+ой#1+2+ами #1+2+е#1+2+ах	239	0=арка, 1=ар, 2=к
8	1#1+и#1+а#1+ов#1+у#1+ам #1#1+и#1+ом#1+ами#1+е#1+ах	222	0=аналог, 1=аналог
9	1#1+и#1+а#1+ов#1+у#1+ам #1+а#1+ов#1+ом#1+ами#1+е#1+ах	174	0=академик, 1=академик
10	1+о#1+а#1+а#1#1+у#1+ам #1+о#1+а#1+ом#1+ами#1+е#1+ах	143	0=агентство, 1=агентств

Tab. 12. Most frequent abstract paradigms for Russian verbs

№	Abstract paradigm	Count	Example
1	1+ть#1+ю#1+ешь#1+ет#1+ем#1+ете#1+ют #1+л#1+ла#1+ло#1+ли#1+й#1+йте	1,316	0=арестовывать, 1=арестовыва
2	1+ться#1+юсь#1+ешься#1+ется#1+емся #1+етесь#1+ются#1+лся#1+лась#1+лось #1+лись#1+йся#1+йтесь	568	0=барахтаться, 1=барахта
3	1+овать#1+ую#1+уешь#1+ует#1+уем #1+уете#1+уют#1+овал#1+овала#1+овало #1+овали#1+уй#1+уйте	302	0=агитировать, 1=агитир
4	1+ить#1+ю#1+ишь#1+ит#1+им#1+ите #1+ят#1+ил#1+ила#1+ило#1+или#1+и #1+ите	192	0=благодарить, 1=благодар
5	1+ить#1+у#1+ишь#1+ит#1+им#1+ите#1+ат #1+ил#1+ила#1+ило#1+или#1+и#1+ите	117	0=вершить, 1=верш
6	1+ить#1+лю#1+ишь#1+ит#1+им#1+ите #1+ят#1+ил#1+ила#1+ило#1+или#1+и #1+ите	116	0=благословить, 1=благослов
7	1+иться#1+юсь#1+ишься#1+ится#1+имся #1+итесь#1+ятся#1+ился#1+илась#1+илось #1+ились#1+ись#1+итесь	104	0=валиться, 1=вал
8	1+дить#1+жу#1+дишь#1+дит#1+дим #1+дите#1+дят#1+дил#1+дила#1+дило #1+дили#1+ди#1+дите	89	0=бродить, 1=бро
9	1+оваться#1+уюсь#1+уешься#1+уется #1+уемся#1+уетесь#1+уются#1+овался #1+овалась#1+овалось#1+овались #1+уйся#1+уйтесь	71	0=адаптироваться, 1=адаптир
10	1+уть#1+у#1+ёшь#1+ёт#1+ём#1+ёте#1+ут #1+ул#1+ула#1+уло#1+ули#1+и#1+ите	66	0=блеснуть, 1=блесн

AUTOMATIC SPELLING CORRECTION FOR RUSSIAN SOCIAL MEDIA TEXTS

Sorokin A. A. (alexey.sorokin@list.ru)[1,2,3],
Shavrina T. O. (rybolos@gmail.com)[1,3]

[1]Lomonosov Moscow State University, Moscow, Russia

[2]Moscow Institute of Science and Technology, Dolgoprudny, Russia

[3]General Internet Corpus of Russian, Moscow, Russia

This paper describes an automatic spelling correction system for Russian. The system utilizes information from different levels, using edit distance for candidate search and a combination of weighted edit distance and language model for candidate hypotheses selection. The hypotheses are then reranked by logistic regression using edit distance score, language model score etc. as features. We also experimented with morphological and semantic features but did not get any advantage. Our system has won the first SpellRuEval competition for Russian spell checkers by all the metrics and achieved F1-Measure of 75%.

Key words: automatic spelling correction, Russian spellchecking, non-word errors, real-word errors, spelling correction, spelling correction for Russian, spelling correction on corpora, social media language

АВТОМАТИЧЕСКОЕ ИСПРАВЛЕНИЕ ОПЕЧАТОК И ОРФОГРАФИЧЕСКИХ ОШИБОК ДЛЯ РУССКОЯЗЫЧНЫХ СОЦИАЛЬНЫХ МЕДИА

Сорокин А. А. (alexey.sorokin@list.ru)[1,2,3],
Шаврина Т. О. (rybolos@gmail.com)[1,3]

[1]МГУ им. М. В. Ломоносова, Москва, Россия;

[2]МФТИ, Долгопрудный, Россия;

[3]ГИКРЯ, Москва, Россия

В данной статье исследуется многоуровневый метод исправления опечаток для русскоязычных текстов, взятых из сети Интернет. Исправление опечаток является особенно важной проблемой в связи с повсеместным использованием социальных медиа в качестве источника для лингвистических исследований. Мы используем комбинацию

нескольких методов, в частности основанных на расстоянии Левенштейна и словарном поиске, а также контекстном ранжировании гипотез с помощью алгоритмов машинного обучения. Наша система заняла 1 место в первом соревновании SpellRuEval по автоматическому исправлению опечаток для русского языка, достигнув F1-меры в 75%.

Ключевые слова: исправление опечаток, автоматическое исправление опечаток, язык социальных медиа, нормализация текста, словарные опечатки

1. Introduction

Spelling correction is one of the oldest and most important problems of computational linguistics. It has attracted many researchers since the pioneer works of Levenshtein and Damerau in the 60-s [Levenshtein, 1965; Damerau, 1964] through the studies on isolated word correction in the beginning of modern NLP era, such as [Kernighan et al., 1989] and early context-based methods [Golding and Roth, 1999] to sophisticated machine-learning based techniques of last decade used in [Whitelaw, 2009] and [Schaback, 2007]. Automatic spellchecking is a problem of high practical importance, especially concerning actual technical requirements of big corpora [Popescu, Phuoc, 2014]. The most straightforward application of it is query correction and completion, used in search engines, as well as orthography correctors which form a part of any modern text editor. More marginal applications include second language learning [Flor, 2012] and grammatical error correction [Rozovskaya, 2013].

In the next section, we describe present-day situation in the field of spelling correction. Section 3 contains a detailed overview of the system developed. Section 4 describes the problems we faced during the development of our system and some open questions left, while Section 5 discusses the results attained. Finally, in Section 6, we offer a brief conclusion of our research and propose some directions for its future application.

2. Past and present of spellchecking: methods and problems

Different researchers focus on different aspects of spelling correction in their studies. Most of the early works dealt with effective search of candidates for typo correction, addressing the problem of fast dictionary lookup and approximate string matching, which is especially important for agglutinative and polysynthetic languages. The task of candidate search is alleviated by the fact that 80% of time a correct word can be obtained from the mistyped word by one primitive edit operation[1]

[1] Primitive edit operations include a) deletion of a single character, b) insertion of a single character, c) substitution of one character for another, d) permutation of a pair of adjacent characters.

[Kukich, 1992]. However, orthographic mistakes usually happen on phonetic level, while spelling correction is performed on the graphic one, which complicates the search when phonetic and graphic representations of do not exactly map to each other: the [Toutanova et al., 2002]. Moreover, not all primitive edit operations and orthographic changes are equiprobable which means that a variant of weighted Levenshtein distance should be used instead of the basic one to achieve better performance [Kernighan et al., 1990; Ristad, Yanilos, 1995].

Correction systems used in text editors, such as Notepad++ or MS Word, usually suggest several candidates, compelling the user to select between them. However, the number of variants even for a medium-length sentence is too high, which implies that ideal spellchecker should be able not only to generate corrections, but also to select the best one in the given context. Another difficulty concerns the typos producing another dictionary word (such as *piece/peace* or *компания/кампания*). Such problems are a subject of context-sensitive spelling correction [Golding, Roth, 1999; Carlson, Fette, 2007]. Most studies address this task in a narrow fashion, trying to correct real-word spelling errors only in the groups of several predefined confusion sets [Pedler, Mitton, 2010]. Then for every confusion set the task becomes a usual classification problem which can be solved using standard machine-learning techniques. However, this method cannot be straightforwardly generalized to real-world spelling correction since the dimension of feature space becomes too high (the most standard features for this task are adjacent words, which means that every pair of dictionary words is a separate feature). Another approach which we pursued in our work is to learn a low-dimensional classifier which uses the scores given by error and language models as its features.

The difficulty of spelling correction also depends from its domain of application and the source language. Indeed, the more fine-grained is the morphology system, the larger is the dictionary, which complicates candidate search and selection. This also makes the data more sparse implying that larger corpora are necessary to learn the language model. Using World Wide Web as a huge unannotated corpora partially solves the problem, but in this case the training data already contains typos which can deteriorate the performance of correction system. Moreover, the percenttage of out-of-vocabulary words, such as proper names, slang and neologisms, is very high for the Web creating another obstacle for the dictionary-based approach. That's why some authors [Whitelaw, 2009] even refuse to use the dictionary basing their algorithms only on corpora frequency.

We are especially interested in spelling correction when applied to social media texts, such as Live Journal, VKontakte and other blogs and social networks. The percentage of misspelled words in such texts is rather high both due to typos and orthography errors and effective correction of such errors is a necessary preliminary condition for further processing such as morphological and syntactical parsing. The percentage of out-of-vocabulary words is also rather high. Our work is a part of General Internet Corpora of Russian (GICR) Project [Belikov et al, 2013]. There are very few works on spelling correction for Russian [Baytin, 2008], [Panina et al., 2013], [Sorokin and Shavrina, 2015]; moreover, the first two are concerned primarily with correction of mistyped search queries, while the latter addresses only to isolated word correction.

3. Our system

Our system participated in the first competition of spellcheckers for Russian SpellRuEval-2016 and won the first place by all the measures, including precision, recall, F1-measure and percentage of correct sentences. We decided to follow the scheme described in [Schaback, 2007] and [Flor, 2013] by collecting scores from different levels including dictionary model, n-gram language model, a weighted error model and morphological error model and combining them in a single linear classifier. We used the reranking approach [Zhang, 2006] often applied in machine translation [Shen et al., 2006]: the algorithm first created n-best lists of candidate sentences according to the simplest of the models and then reranked these hypotheses using logistic regression classifier. We observed that reranking indeed leads to a consistent gain in performance. Other results were quite surprising for us: we observed that morphological and semantic features does not give any further improvement after applying the error model, which either implies that the features we used are too weak to distinguish good hypotheses from the best ones in comparison with other features or that our model of morphology and semantics was not adequate for this task.

3.1. Multi-level spelling correction

In this section we describe our algorithm of spelling errors correction. The algorithm processes one sentence at a time and consists of two stages. On the first stage we rank the candidate hypotheses according to a baseline model. Then for every candidate correction several scores are calculated. These scores include the number of words corrected, the logarithmic probability of the sentence according to the language model, the logarithmic probability of the source sentence to be obtained from the correction by the error model and several other scores characterizing the adequacy and quality of the corrections. These scores were given to a linear classifier as features and the candidate sentence with the highest score was selected. The weights of the classifier were trained on the development set. We describe all these operations further in the article.

3.2. Candidate generation

When generating the candidate sentences, the first step is to find possible corrections for every word in this sentence. The first part of this list consists of all the words on the edit distance 1 from the source as well as the source word itself, no matter whether it appears in the dictionary or not. We used the list of words from ABBYY Compreno [http://www.abbyy.ru/isearch/compreno/] dictionary, which includes approx. 3.7 million words. We store the dictionary as a prefix tree which allows us to effectively search for the words on the distance d or less. We selected $d=1$ since larger values lead to a drastic expansion of candidate list. The search procedure follows the algorithm described in [Oflazer, 1996] with the heuristics used in [Hulden,

2009], for a more detailed description we refer the reader to [Sorokin and Shavrina, 2015]. We transformed all the words to lower case but preserve the information about the capitalization of source word since the abbreviations written by all capitals and lowercase common words have different probabilities to be mistyped. All the words which contain non-alphabetic characters such as Latin letters or digits were copied to the output without candidate search except certain special cases like (в4ера → вчера).

However, this error model does not capture several frequent patterns, such as *цца → тся/ться* transformation in the verb flexion (*появляцца → появляться, появицца → появится*), which is very popular in Russian Internet slang. We deal with this problem by using an analogue of Metaphone algorithm [Philips, 2000], mapping the sounds to their phonetic classes. We used the following table of classes:

Table 1. Mapping from symbols to phonetic codes

code	Russian letters
1	а, о, ы, у, я
3	и, е, ё, ю, я, э
5	б, п
6	в, ф
7	д, т
8	г, к, х
9	л
10	р
11	м
12	н
13	з, с
14	й
15	щ, ч
16	ж, ш
17	ц

The symbols also affect the code of consequent symbols: all the vowels after the *ь, ъ, щ, ч, й* letters were mapped to class 3, as well as to class 1 after *ш, ж* and *ц*. The signs *ь, ъ* were omitted. To deal with multisymbol sequences we mapped the *тс* sequence possibly with sign letters between them to the class 17, we also omitted *т* after a sibilant and before other consonant (for example, in the word *грустный* and compressed consecutive occurrences of the same code to a single one. To use this algorithm in spelling correction, we extended the list of candidates by all the dictionary words having the same phonetic code as the source word. However, these transformation does not capture all irregular patterns, so we handcoded a list of about 50 transformations such as *ваще → вообще* and *грит → говорит*.

Another frequent error pattern is insertion/deletion of space symbol. When the space is inserted in the middle of the word, it is straightforward to model it by allowing the algorithm to traverse from the terminal node of the dictionary tree to its root and paying the special cost for space insertion on this edge. This modification allows us to recognize all the tuples of dictionary word separated by one or more spaces. An analogous problem arises when considering the deletion/insertion of hyphen (-) symbol in composite words, we solved it by adding '-' to the alphabet. This method cannot handle space deletion since we process the sentence word by word. Therefore we performed the candidate search not only for every single word, but also for the groups of two consecutive words.

Given a sentence of 10 words and 3 dictionary candidates for every word in average, which is more probably an underestimate than an overestimate, we obtain approximately 60,000 candidate sentences for every source sentence, which is obviously

impossible to handle. Therefore the procedures of candidate generation and baseline candidate ranking cannot be separated. We rank the candidates by the sum $C_{lm}+C_{err}$ of language model score C_{lm} and basic error model score C_{err} (the least score is the best). We describe language model in the next subsection and now we explain the basic error model.

Given a source sentence $u_1, ..., u_n$, we generate the candidate hypothesis $v_1, ..., v_m$ by groups, for example (1):

(1) *кто то ищо сделал тоже предположение*
кто-то ещё сделал то же предположение

Here we have 5 groups (*кто то, кто-то*), (*ищо ещё*), (*сделал, сделал*), (*тоже, то же*), (*предположение, предположение*) and denote them by g_1, ..., g_5. If the partition of the source-candidate pair of sentences include r groups g_1, ..., g_r, each group including the source word group s_i and correction word group t_i, then the overall cost of this transformation is $-\sum_{i=1}^{r} \log C(s_i \to t_i)$ where $C(s_i \to t_i)$ is the cost of transforming s_i to t_i. This cost is calculated by the following euristics table. The row of a table correspond to the property of s_i, while the column describes the way to obtain t_i from s_i. When a source group is fixed, the weights of different hypotheses are taken from the same row, therefore the weights do not need to sum to 1 since the normalizing coefficient after calculating the logarithm yields a constant summand, which is the same form all candidate word groups. The weights were obtained empirically from the development set by calculating the frequencies of different typo-correction transformations.

Table 2. Weights of different word transformations

	$s_i = t_i$	Levenshtein	Phonetic code	2 words from 1	2 words from 1
no capitals, dictionary word	1	0.005	0.0005	0.001	0.005
initial capital, dictionary word	1	0.001	0.0001	0.0001	0.0001
all capitals, dictionary word	1	0.001	0.0001	0.0001	0.0001
no capitals, dictionary word	0.15	0.6	0.09	0.15	0.01
initial capital, dictionary word	1	0.05	0.01	0.005	0.005
all capitals, dictionary word	1	0.1	0.01	0.01	0.01

3.3. Language model and generation of candidate sentences

Since basic error model score cannot distinguish between different dictionary words on the same Levenshtein distance, we also took into account the language model to obtain the baseline candidate score $C_{lm}+C_{err}$. Provided k is the order of the language model, the language model score of the candidate sentence $t_1, ..., t_m$ equals

$$\begin{aligned} C_{lm} &= -\log p(t_1, \dots, t_m) = -\log(p(t_1)\, p(t_2|t_1) \dots p(t_k|t_1 \dots t_{k-1}) \dots p(t_m|\, t_{m-k+1} \dots t_{m-1})\,) \\ &= -(\log p(t_1) + \log p(t_2|t_1) + \dots + log\, p(t_m|t_{m-k+1} \dots t_{m-1}) \\ &= \textstyle\sum_{i=1}^{i=m} -\log p(t_i|t_{i-k+1} \dots t_{i-1}) \end{aligned}$$

The logarithmic cost in the language model appears to be additive as well as the error model cost. It permits us to apply beam search for pruning partial hypotheses space. Agenda consists of n+1 hypotheses lists, where i-th list store partial hypotheses after processing i words of the sentence and n is the length of the source sentence. Initial agenda item consists of empty hypotheses with initial cost 0. On i-th step we generate all the candidates for the word s_i to expand the hypotheses on the step i−1, as well as the candidates for the group (s_{i-1}, s_i) to expand the hypotheses from the (i−2)-th item of the agenda. For every partial hypothesis we store its current score and the state of the language model (roughly speaking, last (k−1) words). Using this information, we are able to recalculate the score and the state for the expanded hypotheses. We arrange the list items by the states of language models, storing all the partial hypotheses with the same state together. To prune the hypotheses space we preserve only such hypotheses whose score is not greater than the score of the best hypothesis times some constant α.

3.4. Learning of the reranking model

After the previous step we have a list of candidates together with their baseline scores. Now our task is to rerank these candidates using algorithms of machine learning. For this goal we use a linear classifier and determine the best correction using $\hat{c}$the rule

$$\hat{c} = argmax_{c \in C} \sum_i w_i f_i(c) + w_0$$

where C is the set of candidate sentences and f_i are features which we specify below. To learn the weights of the classifier we observe that maximum does not depend on the additive term w_0 therefore only the linear coefficients w_i should be learnt. In machine translation literature the usual approach is to learn these coefficients from the ranking of train hypotheses, however, in our disposal are only the corrections for the training sentences. However, it is sufficient to learn the weights: a good decision function should rank the correct hypothesis higher than the incorrect ones, therefore we have $\sum w_i f_i(\hat{c}) \geq \sum w_i f_i(c)$ for any other hypothesis c. Equivalently, we have $\sum w_i(f_i(\hat{c}) - f_i(c)) \geq 0$. Then our task is to find a linear classifier such that all the vectors of the form $[f_1(\hat{c}) - f_1(c), \dots, f_m(\hat{c}) - f_m(c)\,]$ belong to the positive class and the opposite vectors—to negative. Then our problem is reformulated as usual linear classification problem and can be solved by any of standard algorithms, such as SVM or logistic regression.

In our experiments we used the following list of features (Table 3). When a feature is defined for a single word (say, its capitalization), it means that we sum its values for all the words in the. For example, the unit feature for a single word yields the number of words for the whole sentence.

Table 3. Features used in classification

Feature name	Description
F1	Sentence length in words
F2	Error score C_{err}
F3	Language model score C_{lm}
F4	Number of corrected words
F5	Number of OOV words
F6	Number of corrections in OOV words
F7	Number of corrections in dictionary words
F8	Number of corrections in capitalized words
F9	Number of corrections on edit distance 1
F10	Number of corrections by phonetic similarity
F11	Number of corrections by word lists
F12	Number of 1 → 2 corrections (space insertions)
F13	Number of 2 → 1 corrections (space deletions)
F14	Number of OOV words having dictionary partitions
F15	Morphological model score
F16	Weighted edit distance score
F17, F18	Semantic model score
F19, F20	Prepositional model score

The calculation of features F1–F14 is straightforward: we just memorize their values for single words in the candidate generation phase and sum these values to obtain the aggregate score. Morphological model score is calculated just as usual language model score, except the n-gram model is built on POS tags instead of words. To learn the weights in the edit distance we use the algorithm of [Brill, Moore, 2000]: we align each word in the development set with its correction and extract all the groups of up to 3 alignment tokens. The only refinement we made is that a token containing a space symbol on either of its sides is not joined to any longer group.

The features F17–F20 were not used in the system we submitted for evaluation since they did not improve performance but we describe them for future research. We tried to use cooccurence information in order to grasp semantic relations. For example nothing but semantics can force the system to prefer the correction (2) "*мне снится, что мы в ссоре и ты на меня ругаешься и сердишься*" instead of (3) "*не снится, что мы в море и ты на меня ругаешься и сердишься*" for the source sentence *мне снится, что мы в соре и ты на меня ругаешься и сердишься* (note that the source word is also in the dictionary). To calculate the semantic score we collected a frequency list for dictionary lemmata from a supplementary corpora

(by frequency we mean the number of sentences containing the lemma). Then we remove the words occurring more than in 1% of sentences (they are noninformative stopwords). We retain 10,000 most frequent lemmata after this removal and for every such lemma collect the list of lemmata cooccurring with it more than a limited number (say, 10) of times. So, for every word from the list we obtain its potential collocations. Then to calculate the semantic score of the sentence we take as features both the number of words from the list of lemmata occurring in the sentence and the number of collocation pairs between these words.

The aim of the prepositional score is to determine the case of a noun knowing a preposition before it. It is often useful because most of the case flections are on the distance of one edit from each other and often simultaneously appear in the candidate list. When the noun immediately follows the preposition it can be captured by a language model, however, often there are intermediate adjectives or dependent noun phrases between the preposition and the noun. To measure this characteristic we collect the total number of prepositions in the sentence, as well as the number of prepositions which do not have nouns or pronouns of the corresponding case to its right. In future research we plan to use several analogous features, characterizing sentence morphology, such as number of coordinated adjective-noun pairs, subject-verb pairs (using gender and number agreement), as well as the total number of nominative case words and finite verbs in the sentence. However, these features are too noisy when collected from a corpus without morphological disambiguation and we do not have access to disambiguated corpora of sufficient size. Since straightforward addition of such features did not improve performance and even led to slight degradation, we decided not to use them. We rejected from application of morphological and syntactic parsers since there quality on social media texts is moderate especially when these texts contain typos. Therefore the exact role of morphology and semantics in Russian spelling correction is left for future research.

4. Evaluating the system

We tested our system in SpellRuEval-2016 competition of spelling correctors for Russian social media. The development set of the competition consisted of 2,000 sentences from Russian social media texts together with their corrections. The test set included 100,000 sentences only 2,000 of which were used for testing. We used the development set to tune the parameters of baseline error model (see previous section) used in candidate selection as well as to tune the weighted edit distance. To avoid zero probabilities in Brill-Moore method we added 0.1 to the counts of every symbol-to-symbol, symbol-to-space, symbol-to-nothing and nothing-to-symbol corrections, as well as to the counts of each transposition of symbols. Since phonetic similarity corrections such as *ться* → *цца* have a high cost in Levenshtein model which leads to noise and outliers in training data, we bound the obtained weighted distance by a fixed number from above. To train the language model we used a supplementary corpus of 5,000,000 sentences (50,000,000 words) obtained from a sample of GICR. Since these sample contained lemmata and morpho-tags (though only POS tags may

be considered as reliable), we also used it to train morphological and semantic models. Our algorithm was implemented in Python language, we used the KenLM toolkit [Heafield et al., 2013] to train the language model and the realization of logistic regression from scikit-learn [Pedregosa et al., 2011] as a linear classifier.

We report the results both for development and test sets. In the development phase we used one half of the set for training and another one for testing. The baseline model in the table before is the model used before the training phase, in the weighted baseline model we train a linear classifier on three features: the number of words, the error model score and the language model score. The Levenshtein model adds as a feature the weighted edit distance, the competition model also uses features F4–F14 from Table 4 and the morphological model also uses the score of the POS-tag n-grams model.

Table 4. Spellchecking quality on development set

Model	Precision	Recall	F1-measure	Sentence accuracy
Baseline	71.68	78.01	74.71	70.70
Weighted baseline	82.83	77.89	80.28	79.40
Levenshtein	87.69	79.15	83.20	81.90
Competition	88.43	81.44	84.79	83.20
Competition+Morpho	88.15	81.79	84.85	83.00

Table 5. Spellchecking quality on test set

Model	Precision	Recall	F1-measure	Sentence accuracy
Baseline	63.11	67.26	65.12	60.06
Weighted baseline	75.55	64.27	69.46	66.93
Levenshtein	80.58	65.94	72.53	68.63
Competition	81.98	69.25	75.07	70.32
Competition+Morpho	81.12	68.98	74.56	70.22

Our system won the first place among 7 participants by all the measures: precision, recall, F-measure and accuracy (percentage of correctly recovered sentences). Moreover, already the baseline model is on the par with the system on the second place. We observe that learning weights of different components of baseline model indeed improves its performance consistently, as well as replacing standard edit distance by its weighted version in the error model. Using additional features also improves performance quality by several percents. However, enriching the model with morphological features does not affect performance on the development set and leads to slight degradation on the test set. Note that all the systems except the baseline have higher precision than recall.

5. Results and discussion

Analyzing the mistakes, we have found two main sources of them: the first are space/hyphen errors and the second—real-word errors. For example, in all the 7 cases when the word "еслиб" should be corrected to "*если б*" (for example, in (4) "*страшно представить еслиб с ней что-то случилось*"), it was erroneously replaced by "*если*". Note that this typo is indeed rather difficult: the system suggestion is also a grammatically correct sentence so language model cannot resolve this ambiguity. Moreover, every word sequence that can follow "*если б*", can potentially follow *если* as well. There only part of our system which can capture such cases is the weighted edit distance model, but its influence is overweighed by other factors.

The second source of errors are real-word errors. In many cases they are in fact grammatical errors like (5) "*у этой девченки одни плюсы и не одного минуса*". Sometimes the system cannot select a correct word between two members of a confusion set such as "*формации/фармации*" in (6) "*если говорить точно то эти две фармации исторически противостоящие есть свойства одного*". Though the trigram *и ни одного* is more frequent than its counterpart *и не одного*, the cost of correction in a dictionary word *не* appears to be too high. Real-word errors of the second type could be potentially resolved using cooccurence statistics ("*формации*" is more likely to appear together with "*исторически*" than the other variant), probably, using larger corpus to train the language model or more powerful semantic representation like Word2Vec could help in this case. A minority of errors is also due to incomplete list of informal variations like сення/сегодня or using wrong wordform of a correct lexeme like in (7) "*мне было очень страшно казалось что по дороге нам встретиться или тигр или егеря или бандиты*". We plan to deal with grammar errors and real-word errors in a separate study. It is not an obvious question, whether they can be resolved without any handcoding of grammar and morphology rules.

Table 6. Final quality attained

scientific group, SpellRuEval-2016	Precision	Recall	F-measure	Accuracy	Place
GICR corpora, MSU	81.98	69.25	75.07	70.32	1 of 7

There is still a large room to improve our current model. First of all, results of [Schaback, 2007] and [Flor, 2012] demonstrate that proper usage of morphology and semantics consistently ameliorates performance, which is not the case for our system. It means that POS tags alone do not carry enough information for reliable disambiguation and more subtle morphological categories should be taken into account. As we have already said, this hypothesis should be tested on high-quality morphologically annotated corpus of sufficiently large size. Our current model of semantics representation is one of the simplest ones, therefore only usage of more fine-grained one could resolve the question, whether semantic and morphological information could be helpful for Russian social media.

6. Conclusion

We have developed a system for automatic spelling correction for Russian social media texts. We have tested it in the competition of spellcheckers SpellRuEval during Dialogue Evaluation-2016, where our system won the first place by all the metrics, reaching the F1-Measure of 75%. We used edit distance together with phonetic similarity to select correction candidates, language model together with error model to score these candidates and linear classification algorithms to rerank them. The features used in the last stage include error model score, weighted Levenshtein distance between the candidate and correction, language model score and several other features like number of corrections in dictionary and non-dictionary words, capitalization, etc. The most straightforward way to improve our system is to use linguistically-oriented features like morphology, cooccurence and collocation scores, grammatical correctness of the sentence and so on. Since our system is rather simple, we hope it could serve as a baseline for future Russian spelling correction systems. We think it could also be useful in similar tasks like grammar correction or normalization of social media texts. The system can also be successfully applied on big data collections from Russian Web, and is likely to become a part of NLP-tools used on GICR—this gives other researchers the advantages of having corpus with more diverse automatic annotation and better POS-tagging and lemmatization.

7. Acknowledgements

Authors of this paper express their sincere thanks to all the students that helped in the annotation of the training set. We also thank Elena Rykunova for her intense assistance in technological support of the web forms and databases for making the golden standard.

References

1. *Bajtin A.* (2008), Search query correction in Yandex [Ispravlenie poiskovykh zaprosov v Yandekse], Russian Internet technologies [Rossijskie Internet-tekhnologii], 2008.
2. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.,* (2013), Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. Proceedings of Web as Corpus Workshop (WAC-8), Lancaster.
3. *Brill E., Moore R. C.* (2000) An improved error model for noisy channel spelling correction. In ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, p. 286–293. Association for Computational Linguistics
4. *Damerau F. J.* (1964) A technique for computer detection and correction of spelling errors. Communications of the Association for Computing Machinery, Vol. 7, No. 3, pp. 171–176.

5. *Flor M.* (2012) Four types of context for automatic spelling correction //TAL.—Vol. 53.—Vol. 3.—pp. 61–99.
6. *Golding A. R., Roth D.* (1999) A winnow-based approach to context-sensitive spelling correction //Machine learning.—Vol. 34.—№ 1–3.—pp. 107–130.
7. *Heafield K., Pouzyrevsky I., Clark J., Koehn I.* (2013) Scalable Modified Kneser-Ney Language Model Estimation //ACL (2).—p. 690–696.
8. *Huldén, M.* (2009) Fast approximate string matching with finite automata. // Procesamiento del lenguaje natural, Vol. 43, pp. 57–64.
9. *Kukich, K.* (1992) Techniques for automatically correcting words in texts. ACM Computing Surveys 24, pp. 377–439.
10. *Kernighan M. D., Church K. W., and Gale W. A.* (1990) A spelling correction program based on a noisy channel model. In Proceedings of the 13th conference on Computational linguistics, pp. 205–210. Association for Computational Linguistics.
11. *Levenshtein V. A.* (1965) Binary codes capable of correcting deletions, insertions and reversals [Dvoichnye kody s ispravleniem udalenij, vstavok i zamen simvolov], Doklady of the Soviet Academy of Sciences [Doklady Akademij Nauk SSSR], 1965, Vol. 163, No. 4, pp. 845–848.
12. *Loukashevich N. V., Dobrov B. V.* (2006) Ontologies for natural language processing: description of concepts and lexical senses. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2006”
13. *Mark D. Kernighan, Kenneth W. Church, and William A. Gale.* (1990) A spelling correction program based on a noisy channel model. In Proceedings of the 13th conference on Computational linguistics, pp. 205–210. Association for Computational Linguistics.
14. *Oflazer K.* (1996) Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction //Computational Linguistics.—Vol. 22.—Vol. 1.—pp. 73–89.
15. *Panina M. F., Baitin A. V., Galinskaya I. E.* (2013) Context-independent autocorrection of query spelling errors. [Avtomaticheskoe ispravlenie opechatok v poiskovykh zaprosakh bez ucheta konteksta], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2013” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2013”], Bekasovo, pp. 556–568.
16. *Pedler J., Mitton R.* (2010). A large list of confusion sets for spellchecking assessed against a corpus of real-word errors //Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC’10).
17. *Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.* (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
18. *Philips L.* (2000) The double metaphone search algorithm // C/C++ users journal.—Vol. 18.—№ 6.—p. 38–43.

19. *Popescu O., Phuoc An Vo N.* (2014) Fast and Accurate Misspelling Correction in Large Corpora. Proceedings of EMNLP 2014: Conference on Empirical Methods in Natural Language Processing. Doha, Qatar.
20. *Ristad E. S., Yianilos P. N.* (1998) Learning string-edit distance //Pattern Analysis and Machine Intelligence, IEEE Transactions on.—Vol. 20.—№ 5.—pp. 522–532.
21. *Rozovskaya A., Roth D.* (2013) Joint learning and inference for grammatical error correction //Urbana.—Vol. 51.—pp. 61–801.
22. *Schaback J., Li F.* (2007) Multi-level feature extraction for spelling correction // IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data.—pp. 79–86.
23. *Shavrina T., Sorokin A.* (2015) Modeling Advanced Lemmatization for Russian Language Using TnT-Russian Morphological Parser. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2015", RSUH, Moscow
24. *Shen L., Sarkar A., Och F. J.* (2004) Discriminative Reranking for Machine Translation //HLT-NAACL.—p. 177–184.
25. *Toutanova K., Moore R. C.* (2002) Pronunciation modeling for improved spelling correction //Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.—Association for Computational Linguistics—pp. 144–151.
26. *Whitelaw C., Hutchinson B., Chung G. Y., Ellis G.* (2009) Using the web for language independent spellchecking and autocorrection. // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2, pp. 890–899. Association for Computational Linguistics.
27. *Zhang, Y., He, P., Xiang, W., & Li, M.* (2006) Discriminative reranking for spelling correction. // Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation—pp. 64–71.

FACTRUEVAL 2016: EVALUATION OF NAMED ENTITY RECOGNITION AND FACT EXTRACTION SYSTEMS FOR RUSSIAN

Starostin A. S. (astarostin@abbyy.com)[1],
Bocharov V. V. (bocharov@opencorpora.org)[2,3],
Alexeeva S. V. (sv.bichineva@gmail.com)[2,3],
Bodrova A. A. (anastasie.bodrova@gmail.com)[2,3],
Chuchunkov A. S. (alex.chuchunkov@gmail.com)[2],
Dzhumaev S. S. (sdzhumaev@abbyy.com)[1],
Efimenko I. V. (veassi@mail.ru)[4],
Granovsky D. V. (dima.granovsky@gmail.com)[2],
Khoroshevsky V. F. (v.khor@mail.ru)[5],
Krylova I. V. (krylova93@gmail.com)[2],
Nikolaeva M. A. (mary.nikolaeva@gmail.com)[2],
Smurov I. M. (ismurov@abbyy.com)[1],
Toldova S. Y. (stoldova@hse.ru)[6]

[1]ABBYY, Moscow, Russia
[2]OpenCorpora.org
[3]St. Petersburg State University
[4]Semantic Hub, Moscow, Russia
[5]Dorodnicvn Computing Centre, RAS (CC RAS)
[6]Russian Research University—Higher School of Economics

In this paper, we describe the rules and results of the FactRuEval information extraction competition held in 2016 as part of the Dialogue Evaluation initiative in the run-up to Dialogue 2016. The systems were to extract information from Russian texts and competed in two named entity extraction tracks and one fact extraction track. The paper describes the tasks set before the participants and presents the scores achieved by the contending systems. Additionally, we dwell upon the scoring methods employed for evaluating the results of all the three tracks and provide some preliminary analysis of the state of the art in Information Extraction for Russian texts. We also provide a detailed description of the composition and general organization of the annotated corpus created for the competition by volunteers using the OpenCorpora.org platform. The corpus is publicly available and is expected to evolve in the future.

Key words: information extraction, evaluation, named entity recognition, fact extraction, relation extraction

FACTRUEVAL 2016: ТЕСТИРОВАНИЕ СИСТЕМ ВЫДЕЛЕНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ И ФАКТОВ ДЛЯ РУССКОГО ЯЗЫКА

Старостин А. С. (astarostin@abbyy.com)[1],
Бочаров В. В. (bocharov@opencorpora.org)[2,3],
Алексеева С. В. (sv.bichineva@gmail.com)[2,3],
Бодрова А. А. (anastasie.bodrova@gmail.com)[2,3],
Чучунков А. С. (alex.chuchunkov@gmail.com)[2],
Джумаев С. С. (sdzhumaev@abbyy.com)[1],
Ефименко И. В. (veassi@mail.ru)[4],
Грановский Д. В. (dima.granovsky@gmail.com)[2],
Хорошевский В. Ф. (v.khor@mail.ru)[5],
Крылова И. В. (krylova93@gmail.com)[2],
Николаева М. А. (mary.nikolaeva@gmail.com)[2],
Смуров И. М. (ismurov@abbyy.com)[1],
Толдова С. Ю. (stoldova@hse.ru)[6]

[1]ABBYY, Москва, Россия
[2]Проект «Открытый Корпус»
[3]Санкт-Петербургский государственный университет
[4]Semantic Hub, Москва, Россия
[5]Вычислительный центр им. А. А. Дородницына РАН
[6]Высшая школа экономики, Москва, Россия

Статья описывает правила и результаты соревнования по извлечению информации из русскоязычных текстов FactRuEval, проводившегося в 2016 г. в рамках инициативы Dialogue Evaluation, приуроченной к конференции Диалог 2016. Соревнование включало две дорожки по извлечению именованных сущностей и одну дорожку по извлечению фактов. В работе излагаются задачи, ставившиеся перед участниками соревнования, и приводятся оценки качества работы систем участников. Кроме того, обсуждаются методы оценки качества для всех трех дорожек и делаются предварительные выводы о современном положении дел в области извлечения информации для русского языка. Особое внимание в работе уделяется составу и устройству размеченного корпуса, созданного в процессе проведения соревнования усилиями волонтеров в рамках платформы OpenCorpora.org. Этот корпус является общедоступным и его планируется развивать в дальнейшем.

Ключевые слова: извлечение информации, тестирование систем, выделение именованных сущностей, выделение фактов, выделение отношений

1. Introduction

At the beginning of 2016, a competition was held for systems capable of extracting information from Russian texts. The competition was given the name FactRuEval and was part of the Dialogue Evaluation initiative. This paper is a report on the results of the competition. As organizers of the competition, we set ourselves three main goals:

- Create an infrastructure for regular evaluation of information extraction systems
- Hold the first evaluation event to analyze state-of-the-art Russian-language information extraction systems
- Create a publicly available corpus that can be used for evaluation and further development of information extraction systems

All of the above goals were successfully achieved. Using the OpenCorpora.org platform, we created a technology for annotating corpora geared towards information extraction needs. A generalized annotation model was developed, which was then used to create a gold-standard markup for several classes of information extraction tasks. Additionally, comparator software was developed, enabling automated comparisons of test markups with the gold standard. Both the annotation model and the comparator software can be used for future evaluations with multiple domain-specific tracks.

Using the evaluation infrastructure mentioned above we held three tracks—one involved the classic task of named entity recognition in the tradition of MUC (see, for example, [Grishman and Sundheim, 1996]) and CoNLL (see, for example, [Tjong Kim Sang and De Meulder, 2003]), another evaluated recognition of named entities with attributes[1], and the third was a competition of fact extracting systems. In the "Tracks" section below, we provide a brief description of the tasks set before the participants in each of the three tracks. A total of thirteen systems were enrolled, some of them being commercial software and some developed for research purposes. Since the competition was anonymous, the "Participants" section contains only general information about the systems without giving their exact names. The "Results" section lists the scores achieved by the competing systems. These scores were obtained using the comparator software mentioned above. The principles underlying the comparator tool are described in the "Evaluation Methods" section.

An annotated corpus that was used during competition contains a demo (or training) subcorpus and a test subcorpus. Both of them were annotated by volunteers using the OpenCorpora.org platform. The "Corpus and Markup" section describes the generalized model used by the annotators, provides some statistics, and discusses the handling of disagreements that arose among the annotators when dealing with some entities. The corpus is publicly available for download and use. Plans for expanding and improving the corpus are also discussed in "Corpus and Markup" section.

---

[1] By "named entity with attributes recognition" we mean recognition of the entity mentions along with specification of simple string values of particular attributes (for example, surname or first name for the Person entity). This task differs strongly from the relation or fact extraction task because entity attributes may have only string values. Moreover such values are always based on some internal parts of the whole entity mention.

Due to space constraints, we have been unable to provide complete details in some of the sections. Readers requiring more information are invited to visit https://github.com/dialogue-evaluation/factRuEval-2016, where they will find all the competition materials.

2. Related Work

The organization of the competition was based on the experience of the international evaluation events devoted to Named entities recognition (NER), relation detection and fact extraction tasks. The first evaluation event devoted to these tasks was inspired by DARPA and was held at Message Understanding Conference (MUC) in 1987–1997 (see, for example, [Grishman and Sundheim, 1996]). Initially, information extraction tasks focused on military messages and information concerning terrorist activities. Later, the focus shifted to newswire articles, from which not only military but also economic information was extracted. It was at the MUC events that the first evaluation principles were laid down and guidelines were developed for the creation of gold-standard annotated corpora enabling comparisons of different information extraction system. Starting from 1999, these tracks in the evaluation events have become part of the Automatic Content Extraction (ACE) program. Detailed descriptions of the tasks, data, and rules over the years are available at https://www.ldc.upenn.edu/collaborations/past-proiects/ace/annotation-tasks-and-specifications (see, for example, [Doddington et al., 2004]). The ACE datasets have included not only English texts, but also texts in Arabic and Chinese. Similar tasks have also been set by CoNLL (see, for example, [Tjong Kim Sang and De Meulder, 2003]).

From 2009 onwards, named entity, relation, and fact extraction tasks were also set in the Knowledge Base Population section of the Text Analysis Conference (TAC), http://www.nist.gov/tac/publications/ (see, for example, [Surdeanu 2013]). The TAC tasks are formulated somewhat differently from ACE's and require some transformation of extracted information into structured data. Contending systems have either to populate a database with information about the objects they detect in the texts and relations among them (Cold Start KBP), or link mentions in the texts to the relevant objects in a database. The task is known as Tri-Lingual Entity Discovery and Linking. There are also relation detection and event extraction tracks in TAC competitions.

Evaluation methods employed within different contests vary from the simple procedure used by CoNLL to the complicated methods employed by ACE. Under the CoNLL rules, only the exact matches between the text fragments detected by a system and those in the gold standard are considered as true positives, with false positives not penalized. Given the lack of gold-standard corpora for Russian and considerable variation in the standards adopted for the existing Russian-language systems, we decided against using this direct evaluation method in the FactRuEval competition.

ACE uses a more sophisticated algorithm [NIST: ACE08 Evaluation Plan], with different named entities assigned different weights, making the interpretation and comparison of results complicated [Nadeau and Sekine, 2007]. The evaluation method used by FactRuEval is largely similar to that used by ACE, allowing one and the same entity to be marked up in several different ways. Scores are computed by assessing

how close a participant's markup is to that in the gold standard (see the "Evaluation Methods" section for details).

We also took into consideration the experience of evaluation events for named entities recognition and relation and fact extraction for some particular languages. For example, such an event (EVALITA (http://www.evalita.it, see, for example, [Caselli et al., 2014]) is held annually for Italian.

In Russia, the first competitions for fact detection systems were held from 2004 to 2006 as part of the ROMIP workshop (http://romip.ru/). The ROMIP tasks in 2005 and 2006 involved named entity recognition and fact extraction (employers' names, ownership of an organization, see [Nekrestjanov and Nekrestjanova, 2006]). The systems had to select text fragments where a certain event was mentioned. However, the ROMIP events attracted only a small number of participants (with only two systems competing in the 2006 fact detection track). When developing the tasks for the 2016 FactRuEval competition, we also drew on the experience gained from the relevant ROMIP tracks.

The number of systems working with texts in Russian has grown considerably in the past few years. Starting from 2010, a number of RU-EVAL events have been held, where certain automated text processing modules are evaluated, including morphology and syntax modules (see [Toldova et al., 2015], http://ru-eval.ru/new/). The RU-EVAL events have provided an insight into the current state of the art in automated processing of Russian texts at the basic levels of linguistic analysis and resulted in the creation of datasets on which text processing systems can be tested.

Being focused on information extraction tasks, FactRuEval is the next step in the direction provided by RU-EVAL. Due to a lack of generally agreed on standards for detecting named entities and facts in Russian texts, we opted for more flexible annotation and evaluation principles similar to those used by ACE. We developed a software tool for creating corpora with sophisticated annotation and organized annotation work on a gold-standard corpus featuring the very basic entities. Following in the footsteps of CoNLL, we also developed and made publicly available a software tool that automatically compares the results obtained by competing systems against the gold standard.

3. Tracks

There were three tracks in FactRuEval, with the systems competing in named entities recognition, extracting entities with attributes, and extracting facts. Participants could enroll their systems in any of the tracks. The rules for each track are described in detail in a separate document which is available on the FactRuEval website, so here we provide only a brief overview.

3.1. Named entity recognition track

This track posed the classic task of Named Entity Recognition [Grishman and Sundheim, 1996]. The competing systems had to locate the mentions of entities of particular types. Entities of the following three types had to be recognized:

1. Person
2. Organization
3. Location (place names)[2]

3.2. Named entities with attributes recognition track

In this track, the participants had to list unique named entities of specific types detected in the texts. For each entity, certain string fields had to be filled with normalized values if the corresponding information was available in the texts.

Normalization required transforming phrases to their canonical form. In most cases, this meant recasting the corresponding text fragment in the nominative case (preserving the grammatical agreement between the elements).

The types of entities in this track were identical to those in the entity extraction track[3].

An essential requirement was that the lists of named entities generated by the systems should not contain any duplicates. By prohibiting duplicates we could evaluate the systems' ability to locally identify the referents of named entities, which has important practical applications. When scoring the results in this track, both inclusion of duplicates ("underidentification") and collapsing different entities into one ("overidentification") were penalized.

3.3. Fact extraction track

In this track, the participants had to detect facts of specific types in the texts. A fact is a relation between several entities (a mention of a certain type of situation, with participants playing certain roles). Only those facts had to be extracted which were *explicitly* mentioned in the texts. The fact fields had to be filled with string values allowing unambiguous identification of the corresponding named entities. Facts of the following types had to be extracted:

- Occupation (employment of a person by an organization[4])
- Deal (some interaction of economic nature between people or organizations)[5]

---

[2] In one variation of the track, the participants had to distinguish between normal mentions of locations and mentions of locations in what may be termed "organization uses" (for example, *"Russia announces counter-sanctions"*). A similar class of entities was given the tag GPE in CoNLL competitions ([Tjong Kim Sang and De Meulder, 2003]).

[3] In this track, organization uses of locations were treated as ordinary locations.

[4] In this track, all contexts that were allowed for organizations were also allowed for "organization uses" of locations.

[5] We had planned to work out a classification of deals while annotating the corpus and give participants the chance to compete in identifying different subtypes of deal. Unfortunately, the selected texts contained too few mentions of deals and we had to give up the idea. For next year's competition we are planning to collect a specialist corpus containing a sufficient number of mentions of various types of deal.

- Ownership (of an organization by a person or organization)
- Meeting (of two or more people)

It is important to note that the competing systems were expected to extract fact instances at text level rather than at sentence level. Different field fillers could be mentioned in different sentences linked by anaphora. For example, from the text fragment *"Russian Milk Ltd. is a highly profitable company. This Friday, it was bought by the famous entrepreneur J. J. Ivanov"* the participants had to extract the fact of a deal between two parties (Russian Milk Ltd. and J. J. Ivanov)[6]. The most complex variation of this track required that the participants distinguish between actual facts[7] and all other kinds of facts (i.e. facts mentioned in negative, future, conditional, etc. contexts).

4. Corpus and Markup

An important objective of the FactRuEval project was to create an open annotated corpus of Russian texts that could be used for future evaluations. To achieve this, we had to develop an annotation model that would cover the main tasks solved by information extraction systems. Such a model was successfully developed and subsequently used to annotate 255 documents. We will first describe the annotation model and then provide some statistics for the annotated corpus.

4.1. Annotation model

The markup has four layers. The first two layers contain annotated mentions of entities, the third layer contains coreference relations, and the fourth layer groups entities into facts. The entity markup (the first two layers) was used to evaluate named entity extraction systems competing in the first two tracks. The results shown in the second track were additionally evaluated using the third markup layer. The fact extraction systems were evaluated using all four markup layers.

As the FactRuEval 2016 tracks sometimes allowed multiple versions of correct or partially correct markup, it was decided to include several markup variants for some of the objects in the test corpus.

4.1.1. Layer 1 markup: spans

In layer 1, typified spans have been marked up in the texts. These are **continuous chains of words** labelled with one or more predefined tags (e.g. "surname", "org_name", "loc_descr"). It is assumed that each type of marked up object has its own set of tags (i.e. types of spans). For example, in the case of people, we had to distinguish

[6] All facts that required anaphora resolution for their successful extraction were marked as "difficult to extract", resulting in two scores for fact extraction—one for extracting easily detectable facts (i.e. those that were stated in their entirety in one sentence) and one for extracting all facts.

[7] Or, to be more precise, facts represented in the texts as having actually occurred.

first names, surnames, patronymics, and nicknames, while in the case of organizations and locations, we had to distinguish object names ("org_name" and "loc_name") and the object descriptors ("org_descr" and "loc_descr")[8].

4.1.2. Layer 2 markup: object mentions

In layer 2, spans are grouped into **object mentions**. Object mentions are also typified. The types of object mentions correspond to the types of entities involved. For example, the following already mentioned types of entity were marked up: people, organizations, locations, and organization uses of locations.

Several mentions may share common spans. This most commonly occurs when annotating coordinated items where two mentions of an object share a common descriptor or where two people are mentioned sharing the same surname (see Fig. 1).

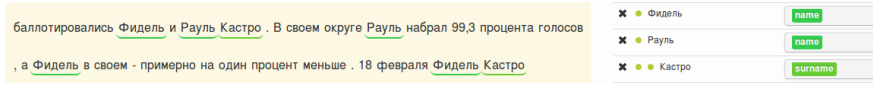


Fig. 1. Two mentions sharing a common descriptor

4.1.3. Layer 3 markup: coreference

In layer 3, object mentions from layer 2 contained in the same text and having the same referent are grouped together (see Fig. 2). Such group is called an identified object. Each group may be linked to an object identifier in an external database (e.g. Wikidata).

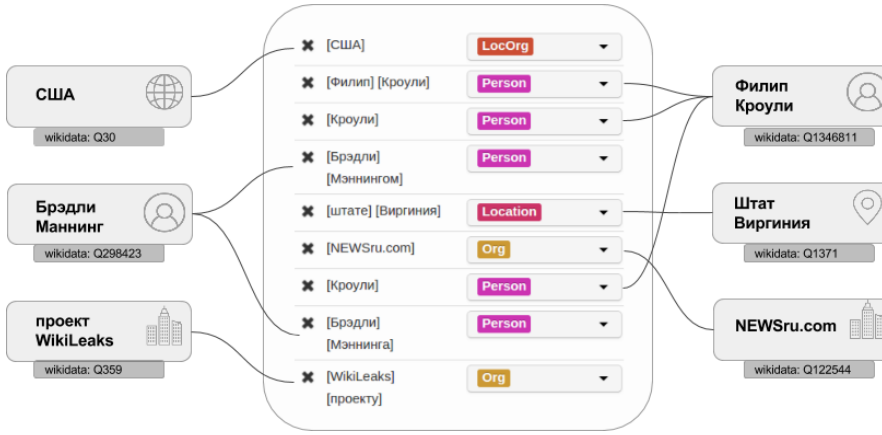


Fig. 2. An example of an identified object

[8] Descriptors are words or word combinations denoting a superordinate concept. For example, "company" is a descriptor in the phrase *"XYZ company"*. Distinguishing between names and descriptors makes it easier to mark up discontinuous mentions of entities. For example, in *"Michurin and Lenin Avenues"* the mention of Michurin Avenue is discontinuous.

4.1.4. Layer 4 markup: facts

A fact is a typified relation between multiple identified objects mentioned in a text. The type of a fact determines what fields it may have. Each field has a name and a list of possible types of object that may fill it. Besides object fields, a fact have string fields. These are filled not by arbitrary strings but by sets of spans (in the general case, by multiple sets of spans, as markup variants are allowed). In a sense, such sets of spans may be considered mentions of virtual objects, i.e. objects that need not be annotated in layers 2 and 3. Fig. 3 illustrates an Occupation fact.

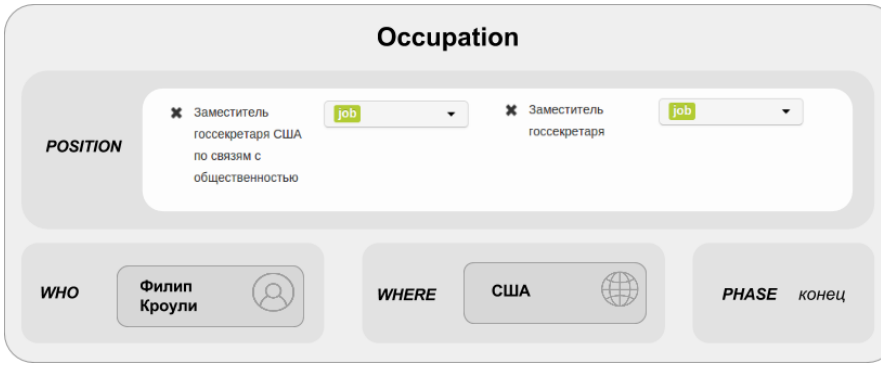


Fig. 3. An example of a fact

It is important that, unlike mentions of objects, facts have no direct links with the original text. They are related to the entire text rather than to a specific text fragment. This approach was adopted because any attempt to link a fact with a specific text fragment often causes disagreement among human annotators and makes it hard to lay down clear requirements for competing systems.

Firstly, it is often the case that a fact is expressed across multiple sentences by means of various anaphoric devices. Secondly, sometimes a fact may logically follow from the text without being explicitly stated[9].

4.2. Corpus characteristics

The FactRuEval corpus consists of newswire and analytical texts in Russian dealing with social and political issues. The texts were gathered from the following sources:

- Private Correspondent (http://www.chaskor.ru/)
- Wikinews (https://ru.wikinews.org)

The corpus was split into two parts—a demo corpus of 122 texts and a test corpus of 133 texts. The demo corpus had been sent out to the participants before the start of the competition. The participants could both test and train their systems on this

[9] This case was excluded from the competition, but obviously we had to keep it in mind when developing the annotation model.

corpus. The test corpus was made available to the participants once the competition ended. During the competition, the participants received a collection of approximately 30,000 documents, which also included documents from the test corpus (of course, the participants did not know which documents came from the test corpus). The text statistics are provided in Table 1. The markup statistics are provided in Table 2.

Table 1. Text statistics

Total texts		**Total characters**		**Total tokens**		**Total sentences**	
Demo Set	**Test Set**	**Demo Set**	**Test Set**	**Demo Set**	**Test Set**	**Demo Set**	**Test Set**
122	133	189,893	460,636	30,940	59,382	1,769	3,138

Table 2. Markup statistics

Spans		**name**		**surname**		**patronymic**		**nickname**		**loc_name**	
Demo	**Test**	Demo	Test	Demo	Test	Demo	Test	Demo	Test	Demo	Test
4,084	**7,670**	531	810	691	1,268	15	27	12	56	1,067	1,367
		loc_descr		org_name		org_descr		job		Other	
		Demo	Test	Demo	Test	Demo	Test	Demo	Test	Demo	Test
		127	194	637	1,628	497	1187	471	915	36	218
Objects		Person		Location		Org		LocOrg		Other	
Demo	**Test**	Demo	Test	Demo	Test	Demo	Test	Demo	Test	Demo	Test
2,611	**5,019**	741	1,388	529	728	787	2,034	553	846	1	23
Facts		Occupation		Deal		Ownership		Meeting			
Demo	**Test**	Demo	Test	Demo	Test	Demo	Test	Demo	Test		
273	**786**	211	444	30	114	17	176	5	51		

The following approach was used to annotate the data for the first two layers. First, the organizers drew up guidelines for annotating each type of entity (these guidelines are available on the FactRuEval website). Next, each paragraph was independently annotated by four volunteer annotators. Any disagreements were resolved through moderation by an expert appointed by the organizers. Layer 2 and 3 annotation was carried out by experts appointed by the organizers. For the demo portion of the corpus, the markup for all the four layers was made available to the participants prior to the competition. The participants had the opportunity (of which they made frequent use) to discuss on the FactRuEval website any problems that they encountered. Following up on these discussions, we made every effort to make the necessary changes to the markup. The markup of the test portion of the corpus was disclosed after the competition ended. The participants had the opportunity to lodge an appeal and state their case for correcting the markup via the FactRuEval website. Some corrections were made to the markup after such appeals, following which the scores were recalculated and finally declared.

The distributed iterative annotation process described above resulted in highly accurate markup. Some problem cases persisted, but for the overwhelming majority of instances a consensus was reached among the annotators, the organizers, and the participants. LocOrg was the most contentious type of entity. There was a lot of debate

as to when it should be treated as an organization and when purely as a location. We had expected LocOrg to cause some controversy, but still decided to have this entity marked up by way of experiment. To account for this uncertainty, we provided a scoring mode in the first track which treated LocOrg as location proper. In this evaluation mode, the three entities have their usual interpretations and the first track is no different from similar evaluations previously conducted in other competitions. The results for the two evaluation modes are provided in the "Results" section.

Unlike the entity markup, the factual markup has certain significant defects, which we hope to rectify before next year's competition. Firstly, too few facts were marked up due to time and labour constraints. In the demo corpus, the fact of employment (Occupation) is the best marked up fact. The other facts are few and far between. The test corpus is slightly better in this respect, but the marked up facts are still far from being representative. Secondly, the demo and test corpora are out of sync in terms of the number of facts they contain. Finally, all facts were marked up by only one (albeit a highly skilful one) annotator. In view of the above, the current factual markup should be treated as a preliminary version which enabled us to carry out a test run of the fact extraction track. For next year's competition, we are planning to resolve these issues and re-run the fact extraction track.

5. Participants

Initially, over sixty teams had expressed their interest in FactRuEval. However, only thirteen actually took part in the competition (not all of them participated in all three tracks).The competition attracted commercial developers, research teams, and news agencies (Fig. 4).

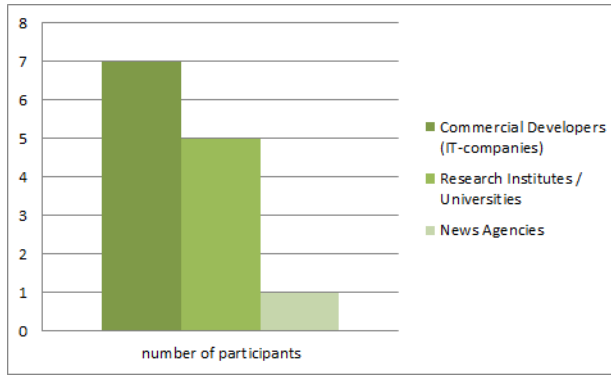


Fig. 4. Types of participants

Many of the participants use the hybrid approach to NLP, i.e. employ both linguistic methods and machine learning. However, when asked by the organizers, most of them described their system either as rule-based or as using machine learning. The rule-based systems prevailed (eight systems were rule-based and five systems used machine learning).

6. Evaluation Methods

This section describes the principles behind the comparator software, which enabled near-automated comparisons of the results. The software was made available to the participants during the competition, allowing them to compare their results against the demo markup. The comparator software is now publicly available. We are planning to improve it and use in future competitions.

The comparator tool examines all possible correspondences between a test markup and the gold-standard markup and chooses the best matches. A good match is a correspondence between objects (i.e. mentions, identified entities or facts) in the gold standard and test markup which meets the following criteria:

- Correspondence must be established between objects of compatible types
- Each object in the test markup must have only one corresponding object in the gold-standard markup
- Each object in the gold standard-markup:
  - Must have only one corresponding object in the test markup (for mentions and entities)
  - May have any number of corresponding objects in the test markup (for facts)

For each pair (or group) of objects, its quality Q, is calculated using a certain formula. Extraction quality of each individual object in the pair (or in the group) is also considered equal $Q(s_i) = Q(t_i) = Q$. The extraction quality of unmatched objects is considered to be zero.

- The obtained values are then used to calculate precision, recall, and F-measure:
- $Precision = \frac{\sum Q(t_i)}{|T'|}, where\ T' = T \backslash T_i$
- $Recall = \frac{\sum Q(s_i)}{|S'|}, where\ S' = S \backslash S_i$
- $F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$

In each track F-measure is chosen as the final score awarded to participants. Below we describe quality estimation principles for each track.

6.1. Entity recognition quality metric

The quality of a matching pair is calculated using this formula:

$$Q(s,t) = \frac{TP}{TP + FP + FN}$$

where

TP is the total number of tokens that belong to both mentions,

FP is the total number of tokens that belong to the test mention, but do not belong to the gold-standard mention,

FN is the total number of tokens that belong to the gold-standard mention, but do not belong to the test mention[10].

6.2. Entity with attributes recognition quality metric

A gold-standard entity *t* may be represented as a set of pairs $< a_i, v_1^i | \ldots | v_{k_i}^i >$ and a test entity *s*—as a set of pairs $<a_i, v_i>$. The quality of a matching pair is calculated using this formula:

$$Q(s,t) = \frac{TP}{TP + FP + FN}$$

where

TP is the number of such pairs $<a, v_1|\ldots|v_N>$ of the gold-standard entity *s* that the entity *t* has at least one pair of type $<a, v_i>$, where $i = 1 \ldots N$,

FN is the number of such pairs $<a, v_1|\ldots|v_N>$ of the gold-standard entity *s* that the entity *t* has no pairs of type $<a, v_i>$, where $i = 1 \ldots N$,

FP is the number of such pairs $<a,v>$ of the test entity *t* that the entity *s* has no pairs of type $<a, v_1|\ldots|v|\ldots|v_N>$[11].

6.3. Fact extraction quality metric

The quality of a group is calculated using two metrics. First, we evaluate the test markup to see how well the system has detected values of fields, using the following metric:

$$ArgQuality(s, \{t_1, \ldots, t_k\}) = JaccardIndex(S.T) = \frac{|S \cap T|}{|S| + |T| - |S \cap T|}$$

where

S is the set of all pairs <field,value > in the gold-standard fact of group (*s*),

T is the set of all pairs < field,value > in all the test facts of group (t_i),

$S \cap T$ is the set of all pairs < field,value > in all the test facts of group (t_i), that have been deemed correct, i.e. those for which a match has been found in *S*.

According to the rules of the third track all values of all fields in the test markup are strings, whereas for the gold-standard markup, all object fields are filled with links to already marked up named entities. To find matches for string fields, we simply look for exact matches[12]. To find matches for object fields, the following rule is used:

---

[10] It should be noted that tokens forming spans of certain types were ignored. For more information, please refer to the instruction manual to the comparator tool, which is available on the FactRuEval website.

[11] To evaluate extraction of entities with attributes, a light comparison mode was available, where FP was always assumed to be 0, i.e. redundant attribute values were not penalized

[12] The only exception is job titles. For them, like for object fields, we allow (i.e. include in the gold standard) variants of normalized names (provided they occur in the text), as well as normalized strings matching exactly a job title mention in the text.

A pair <field R, value X> from the test fact is matched with a pair <field R,link to named entity E> from the gold-standard fact if and only if X is either contained in the allowed names of entity E in the gold-standard markup[13], or exactly (letter by letter, without normalization) matches one of the continuous mentions of the entity in the text.

The second metric shows how well the test markup reflects the co-occurrence of various field values of a fact, or, to put it another way, how well the system has joined together the fragments of the detected facts. To count the quality two graphs are examined, one for the gold-standard and one for the test markup. Let the set of nodes in each graph correspond to the set of correctly detected <field,value> pairs that have been grouped together (see above)—$S \cap T$, $|S \cap T| = n$. Let us assume that a pair of nodes in the graph is linked by an arc if and only if there is an instance of the fact in the corresponding markup which simultaneously contains the <field,value> pairs, corresponding to these nodes.

Obviously, due to the constraints imposed on a group, the graph corresponding to the gold-standard markup will be complete. On the other hand, in the second graph there will appear connected components v_1, ..., v_m of the sizes n_1, ..., n_m, and $\sum n_i = n$. Using the sizes of these connected components, we calculate the second metric as a ratio of the numbers of arcs in the two graphs:

$$IdQuality(s, \{t_1, \dots, t_k\}) = \frac{\sum n_i(n_i - 1)}{n(n-1)}$$

The quality of the entire group is calculated as:

$$Q(s, \{t_1, \dots, t_k\}) = \frac{ArgQuality * (1 + IdQuality)}{2}$$

7. Results

In this section, we present the performance scores of the systems for each of the three tracks. All the participating systems are listed under their code names assigned to them upon enrolling the competition. If a participant sent in more than one run for a track, we give only the scores for the run with the highest F-measure.

7.1. Entity extraction scores

As we mentioned earlier, the results in the first track were scored using two different evaluation modes. The first mode required that the participating systems distinguish between mentions of proper locations and organization uses of locations. The

[13] Here, too, there is one exception. For Occupation facts, the comparator tool automatically adds the names of superordinate organizations (if mentioned in the text) to the allowed names of entities that indicate organizations where people work. For example, if, analyzing the sentence "Ivanov teaches at the Department of History at Moscow University", a system extracts the fact of working at Moscow University rather than at the Department of History at Moscow University, this answer will be treated as correct.

second mode ignored this distinction. The first evaluation mode gave rise to a lot of objections from participants, as there was no generically agreed on definition of "organization use". Despite the fact that a lot of effort had been put into working out a precise definition for this use of locations, some of the participants decided against detecting LocOrg entities. In Table 3 below, we provide scores only for those systems whose output included LocOrg entities. The scores were computed using the evaluation mode that distinguished between proper locations and entities of type LocOrg. The best result (0.809) was shown by the system known as *pink*, with *aquamarine* and *crimson* close behind.

Table 4 lists the scores computed without making a distinction between proper locations and entities of type LocOrg. In this evaluation mode, the best result was shown by the system known to the organizers as *violet*. Very close to *violet* were *pink* and *beige*. A case apart is the system known as *grey*, whose developers only sent in the results for entities of type Person, making comparisons with the other systems difficult. The only thing we can say about *grey* is that it is in the top five systems when it comes to detection of people.

Table 3. Entity extraction scores. Location and LocOrg are treated as two different types of entity

	Overall			Person			Location			Organization			LocOrg		
System	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
aquamarine	0.84	0.78	0.807	0.91	0.92	0.91	0.74	0.70	0.72	0.80	0.73	0.76	0.87	0.67	**0.76**
crimson	0.86	0.74	0.80	0.96	0.88	0.92	0.73	0.60	0.66	0.84	0.69	0.76	0.79	0.72	0.75
orange	0.83	0.74	0.78	0.93	0.87	0.90	0.73	0.69	0.71	0.78	0.66	0.72	0.79	0.73	0.75
pink	0.86	0.76	**0.809**	0.96	0.87	0.91	0.80	0.67	**0.73**	0.86	0.71	0.78	0.74	0.75	0.75
violet	0.79	0.75	0.77	0.94	0.92	**0.93**	0.52	0.86	0.65	0.82	0.76	**0.79**	0.89	0.31	0.46
white	0.74	0.47	0.58	0.95	0.74	0.83	0.43	0.70	0.53	0.87	0.36	0.51	0.06	0.00	0.00

Table 4. Entity extraction. Location and LocOrg are treated as the same type of entity

	Overall			Person			Location			Organization		
System	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Aquamarine	0.88	0.82	0.85	0.91	0.92	0.91	0.96	0.81	0.88	0.80	0.73	0.76
Beige	0.88	0.83	0.86	0.90	0.90	0.90	0.96	0.86	**0.91**	0.81	0.74	0.77
Black	0.86	0.83	0.85	0.91	0.92	0.92	0.96	0.86	0.90	0.74	0.73	0.74
Brown	0.89	0.69	0.78	0.96	0.84	0.90	0.91	0.72	0.80	0.78	0.54	0.64
Crimson	0.92	0.79	0.85	0.96	0.88	0.92	0.96	0.81	0.88	0.84	0.69	0.76
Green	0.90	0.73	0.81	0.93	0.84	0.88	0.95	0.84	0.89	0.82	0.55	0.66
Grey	—	—	—	0.96	0.87	0.91	—	—	—	—	—	—
Orange	0.87	0.78	0.82	0.93	0.87	0.90	0.91	0.84	0.87	0.78	0.66	0.72
Pink	0.92	0.80	0.86	0.96	0.87	0.91	0.94	0.85	0.89	0.86	0.71	0.78
Purple	0.85	0.79	0.82	0.90	0.88	0.89	0.92	0.84	0.88	0.76	0.68	0.71
Ruby	0.88	0.54	0.67	0.92	0.73	0.81	0.89	0.67	0.76	0.78	0.26	0.39
Violet	0.89	0.84	**0.87**	0.94	0.92	**0.93**	0.93	0.87	0.90	0.82	0.76	**0.79**
White	0.93	0.58	0.71	0.95	0.74	0.83	0.93	0.68	0.79	0.87	0.36	0.51

7.2. Scores for extracting entities with attributes

Two evaluation modes were used for this track. The first mode treated redundant attribute values (i.e. those not found in the gold standard) as errors, while the second mode allowed redundancies. The idea behind was to give a fair chance to those systems that made extensive use of encyclopedic information and that, for various reasons, could not remove this information from their output. As it turned out, switching between the two evaluation modes had almost no impact on the scores. Nevertheless, we provide the scores for both modes (Tables 5 and 6). In both cases, the highest F-measure (0.80) was achieved by *pink*. It should be noted, however, that the scores for *violet*, *crimson*, and *aquamarine* were very close to those of *pink*.

Table 5. Scores for extraction of entities with attributes. Redundant attribute values are penalized

	Overall			Person			Location			Organization		
System	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
aquamarine	0.79	0.79	0.79	0.81	0.86	0.83	0.94	0.82	**0.87**	0.68	0.71	0.70
crimson	0.83	0.73	0.78	0.86	0.82	0.84	0.93	0.75	0.83	0.73	0.64	0.68
green	0.76	0.72	0.74	0.79	0.83	0.81	0.90	0.79	0.84	0.64	0.58	0.61
pink	0.84	0.76	**0.80**	0.89	0.82	0.85	0.90	0.81	0.86	0.76	0.66	**0.71**
violet	0.79	0.78	0.78	0.88	0.86	**0.87**	0.84	0.81	0.83	0.68	0.69	0.68

Table 6. Scores for xtraction of entities with attributes. Redundant attribute values are not penalized

	Overall			Person			Location			Organization		
System	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
aquamarine	0.79	0.79	0.79	0.81	0.87	0.84	0.94	0.82	**0.87**	0.68	0.71	0.70
crimson	0.83	0.73	0.78	0.87	0.83	0.85	0.93	0.75	0.83	0.73	0.64	0.68
green	0.77	0.73	0.75	0.80	0.84	0.82	0.90	0.79	0.84	0.64	0.58	0.61
pink	0.86	0.77	**0.81**	0.91	0.84	0.87	0.91	0.82	0.86	0.78	0.68	**0.73**
violet	0.79	0.78	0.79	0.88	0.86	**0.87**	0.84	0.81	0.83	0.68	0.69	0.69

7.3. Fact extraction scores

Only two systems participated in the fact extraction track, *violet* and *green*. We can think of three possible reasons for this lack of participants. Firstly, fact extraction is a very complicated task, with only a handful of teams working on it. Secondly, some of the teams that could potentially participate in the track do not share some of the organizers' ideas (e.g. the idea that facts should be detected at text level). Thirdly, and, perhaps, most importantly, prospective participants were not satisfied with the corpus of facts offered by the organizers (the known problems associated

with the factual aspect of the corpus are described in the "Corpus features" section). The small size of the demo corpus shut out systems that relied on machine learning and made it difficult to fine-tune rule-based systems. We hope that for next year's competition we will have a larger corpus of facts that will attract more participants.

Despite of the shortcomings described above, the fact extraction results obtained by the two participating systems in this year's competition are of some interest. Particularly important are the results of extracting the Occupation fact. The corpus contains a sufficiently large number of instances of this type of fact, so the results shown by *violet*, the winner of this track, are meaningful and can be considered as today's baseline for fact extraction systems working with Russian texts. It would be extremely interesting to analyze the errors made by *violet* to get an insight into what presents the most difficulty to text analysis systems and to have some sort of typology for such problem cases. Additionally, the fact that two systems participated in the fact extraction track and delivered meaningful results shows that the fact markup mechanisms and the comparator tool can be successfully used for future evaluations.

Tables 7 and 8 list the scores awarded for the fact extraction track. The scores in Table 8 were obtained with additional constraints imposed by the comparator tool, penalizing the systems' failure to detect facts expressed using anaphoric devices and detection of "non-facts" mentioned in future tenses or in negative or conditional contexts.

Table 7. Fact extraction scores in standard mode

	Overall			Ownership			Occupation		
System	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
green	0.54	0.23	0.33	0.35	0.09	0.14	0.65	0.36	0.46
violet	0.75	0.38	**0.51**	0.54	0.17	**0.26**	0.80	0.56	**0.66**
				Meeting			Deal		
System				**P**	**R**	**F1**	**P**	**R**	**F1**
green				0.67	0.15	**0.24**	0.18	0.06	0.09
violet				0.87	0.14	0.23	0.68	0.19	**0.30**

Table 8. Fact extraction scores in advanced mode

	Overall			Ownership			Occupation		
System	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
green	0.52	0.22	0.31	0.37	0.09	0.15	0.62	0.30	0.41
violet	0.70	0.36	**0.47**	0.54	0.15	**0.24**	0.74	0.49	**0.59**
				Meeting			Deal		
System				**P**	**R**	**F1**	**P**	**R**	**F1**
green				0.50	0.12	**0.20**	0.15	0.06	0.08
violet				0.58	0.10	0.17	0.63	0.22	**0.32**

8. Conclusion

FactRuEval 2016 gave us some idea of the current capabilities of information extraction systems working with Russian texts. It is clear from the results that the quality of named entity extraction for Russian is comparable to that of systems working with English texts.

We were unable to fully evaluate the quality of fact extraction for two reasons. Firstly, the number of systems that took part in the fact extraction track was too small. Secondly, the corpus offered to the participants did not contain a sufficient number of facts of different types. Of the five types of declared facts, only Occupation had enough marked up instances for meaningful evaluation[14]. Therefore, this year's fact extraction evaluation must be treated as preliminary. Next year we are planning to organize work to gather and mark up a sufficiently large number of fact mentions to have a representative corpus of facts.

An important result of FactRuEval has been the creation of an infrastructure based on the OpenCorpora.org platform that will enable future evaluations of information extraction systems working with Russian texts. We now have a technology in place enabling volunteers to annotate corpora for mentions, objects, and facts. We have also developed a handy comparator tool that compares markups produced by competing systems against the gold standard.

Equally important is the fact that we now have a publicly available gold-standard corpus annotated by volunteers. We hope that its development will continue and invite all interested parties to use this corpus and contribute to it. It would be of great value to all working in information extraction to have various specialized corpora available from OpenCorpora.org containing domain-specific entities and facts from the fields of law, medicine, etc. The existing corpus can also be expanded with averaged and moderated results obtained by the competing systems[15] on 30,000 documents fed to them in the course of the competition.

9. Acknowledgments

We would like to thank all the participants and all those who helped us in organizing FactRuEval 2016. In particular we want to thank Ekaterina Protopopova for her help with gathering documents for markup. We also want to thank her and Liliya Volkova for the careful proofreading of this article.

And of course FactRuEval 2016 would not be possible without all of the volunteers who took part in the annotation effort.

The reported study was partially supported by RFBR, research project No. 15-07-09306 "Evaluation benchmark for information retrieval".

---

[14] And even Occupation instances were often in the form of simple continuous groups of type "*job title-organization-person*".

[15] Only the results from those systems will be used whose developers agreed to such use during registration.

References

1. *Caselli T., Sprugnoli R., Speranza, M., Monachini M.* (2014), EVENTI: EValuation of Events and Temporal INformation at Evalita 2014. Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014 (pp. 27–34). Pisa University Press.
2. *Doddington G. R., Mitchell A., Przybocki M. A., Ramshaw L. A., Strassel S., Weischedel R. M.* (2004), The Automatic Content Extraction (ACE) Program-Tasks, Data and Evaluation, LREC, Vol. 2, p. 1.
3. *Grishman R., Sundheim B.* (1996), Message Understanding Conference-6: A Brief History, COLING, Vol. 96, pp. 466–471.
4. *Nadeau D., Sekine S.* (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3–26.
5. *Nekrestjanov I., Nekrestjanova M.* (2006), ROMIP'2006: Organizers' report. Proceedings on the 4th russian workshop ROMIP'2006 (pp. 7–29), Saint Petersburg.
6. *NIST:* ACE08 Evaluation Plan. http://www.nist.gov/speech/tests/ace/2008/doc/ace08- evalplan.v1.2d.pdf (2008)
7. *Surdeanu M.* (2013), Overview of the TAC2013 knowledge base population evaluation: English slot filling and temporal slot filling, Proceedings of the Sixth Text Analysis Conference (TAC 2013).
8. *Tjong Kim Sang E., De Meulder F.* (2003), Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, pp. 142–147, Association for Computational Linguistics.
9. *Toldova S., Lyashevskaya O., Bonch-Osmolovskaya A. Ionov M.* (2015), Evaluation for morphologically rich language: Russian NLP. Proceedings on the International Conference on Artificial Intelligence (ICAI) (p. 300). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

INFORMATION EXTRACTION BASED ON DEEP SYNTACTIC-SEMANTIC ANALYSIS

Stepanova M. E. (Maria_Ste@abbyy.com)[1],
Budnikov E. A. (Egor_B@abbyy.com)[1],
Chelombeeva A. N. (Antonina_C@abbyy.com)[1],
Matavina P. V. (Polina_Ma@abbyy.com)[1],
Skorinkin D. A. (Daniil_S@abbyy.com)[1,2]

[1]ABBYY, Moscow, Russia
[2]Higher School of Economics, Moscow, Russia

This paper presents a rule-based approach to Information Extraction (IE) task within FactRuEval-2016 competition. Our system is based on ABBYY Compreno Technology. The technology uses the results of deep syntactic-semantic analysis, which leads to significant reduction of the number of necessary rules and makes them laconic.

The evaluation was conducted on FactRuEval dataset. FactRuEval is an open evaluation of IE systems. The participants could take part in three tracks. The first track required to detect the boundaries and type of named entities in a text. The second track required to extract normalized attributes and perform local identification of named entities. The third track required to extract facts of certain types from a text. We took part in all three of the tracks with the nickname *violet.* Our method proved to be successful: we have achieved high F-measures in Named Entity Recognition tracks and the highest F-measure in Fact Extraction track.

Key words: information extraction, named entity recognition, syntactic analysis, anaphora and coreference resolution, fact extraction

ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ НА ОСНОВЕ ГЛУБОКОГО СИНТАКТИКО-СЕМАНТИЧЕСКОГО АНАЛИЗА

Степанова М. Е. (Maria_Ste@abbyy.com)[1],
Будников Е. А. (Egor_B@abbyy.com)[1],
Челомбеева А. Н. (Antonina_C@abbyy.com)[1],
Матавина П. В. (Polina_Ma@abbyy.com)[1],
Скоринкин Д. А. (Daniil_S@abbyy.com)[1, 2]

[1]ABBYY, Москва, Россия
[2]Высшая школа экономики, Москва, Россия

Ключевые слова: извлечение информации, распознавание именованных сущностей, синтаксический анализ, разрешение кореференции, извлечение фактов

1. Introduction

As the quantities of texts available in digital form increase rapidly, there is a growing need for Information Extraction (IE) systems to process them. In recent decades a number of competitions were held to assess performance of such systems for different languages. However, very few attempts were made to evaluate the state of the art for Russian language IE, and until recently no open-access corpora were available for such evaluation. FactRuEval-2016 was intended to amend the situation by running an independent contest between IE systems for Russian language and publishing a freely available corpus afterwards.

For those willing to participate, three independent tracks were provided. The first track tested the standard "baseline" named entity recognition (NER). Track participants were supposed to detect and annotate each entity and correctly attribute its type (Person, Organization, Location, LocOrg) without establishing any coreference chains. The second track involved local identification of entities and extraction of values for their predefined properties/attributes (e.g. name, surname, organization name). The third track evaluated fact extraction. Fact types were limited to four: occupation, business deal, ownership and meeting.

The Information Extraction system we describe in this paper is a model-based one, i.e. it combines rule-based approach, sophisticated language modelling and statistical parsing. The system took part in all three tracks with the results available below (see Results).

2. Related Work

One of the most well-known conferences that addressed the problem of IE was Message Understanding Conference (MUC). The term "Named Entity" was actually coined for the Sixth MUC (Grishman, Sundheim, 1996). ACE, TAC and CoNLL (Tjong Kim Sang, De Meulder, 2003) can be named among other conferences that addressed NER in their tasks. An overview of the main approaches to NER is given in (Nadeau, Sekine, 2007). An Overview of Event Extraction from Text is given in (Hogenboom et al., 2011).

Two main approaches to information extraction are classifier-based and pattern-based. A classifier-based system is, for example, ALICE (Chieu et al., 2003). WHISK (Soderland, 1999) and (Yakushiji et al., 2006) can be named as examples of pattern-based systems.

As for locality of the context being used for event extraction, usually event extraction systems rely on the local context around phrases that are considered as candidates for extraction. Some systems use extraction patterns (Soderland et al., 1995; Riloff, 1996; Yangarber et al., 2000; Califf and Mooney, 2003), which represent the immediate contexts surrounding candidate extractions. Similarly, classifier-based approaches (Freitag, 1998; Freitag and McCallum, 2000; Chieu et al., 2003; Bunescu and Mooney, 2004) rely on features in the immediate context of the candidate extractions.

Patwardhan and Riloff (Patwardhan and Riloff, 2009) introduced an event extraction model that consists of two parts: a sentential event recognition that

determines if a sentence is discussing a domain-relevant event and a role fillers recognizing that identifies phrases as role fillers based upon the assumption that the surrounding context is discussing a relevant event. In our system we use similar approach when we determine events independently from named entities and look for the role fillers in specific semantic slots under a predicate.

Gareev et al. suggested quality baselines for the Russian NER task considering lack of works reporting results for the Russian language (Gareev et al., 2013). The authors also implemented and evaluated two approaches to NER: knowledge-based and statistical. Shelmanov et al. presented and evaluated the pipeline for processing of clinical notes in Russian (Shelmanov et al., 2015). In different tasks they used both rule-based patterns and several supervised machine-learning methods. In (Solovyev et al., 2012) the authors used extraction templates to extract events from texts in Russian. It was noted that information from different sources may be used for building those templates: intuition of the developer, examples from the texts, language model based on Chomsky grammars or other formal language models like in (Mel'čuk, I. A., 1973).

3. Method

3.1. Syntactic-semantic trees

Most of pattern-based systems rest on regular expression patterns. Often the words describing a specific type of event are grouped into semantic classes to reduce the final number of patterns. The change of the domain requires creation of new patterns and semantic classes, which makes the development of such systems resource consuming. Furthermore, the syntactic variability may require creation of a great number of patterns. On the other hand, classifier-based systems tend to have lower recall due to the inability to take into account distant interdependencies and coreference.

To meet these challenges we use a syntactic-semantic parser that allows us to perform full syntactic-semantic analysis of a natural language text (Anisimovich et al., 2012). The syntactic-semantic analysis is based on multilevel natural language model created by linguists and then corpus-trained (Zuev et al., 2013). The output is a forest of dependency-based parse trees augmented with grammatical and semantic information. The trees can be viewed either as projective dependency tree or as constituent tree.

To perform semantic analysis the parser uses semantic hierarchy of language-independent "meanings" (semantic classes). The language-specific lexemes fit into semantic hierarchy as children of semantic classes. The parse tree nodes are augmented with semantic classes as well as the so called semantic slots (semantic roles), this information is language independent. Information on syntactic slots (extended analogue of syntactic functions) and non-tree links (conjunction, pronoun anaphora and other non-local dependencies) is also present. The latter is especially important for the needs of information extraction.

3.2. The information extraction mechanism

The input accepted by the information extraction mechanism is a sequence of syntactic-semantic trees. The output of the extraction mechanism is an RDF graph. The RDF data is consistent with an OWL-DL ontology (W3C, 2004) which is a predefined and static. Information about facts (i.e. situations and events) is modelled in a way that is ideologically similar to that proposed by the W3C consortium for defining N-ary relations (W3C, 2006)

The mechanism of information extraction is controlled by a system of production rules of three types

1. The rules of interpretation of syntactic-semantic structures
2. The rules of identification of information objects
3. Anaphoric rules

The rules are formulated on a special formal language, the syntax of which has the constructions to work with syntactic-semantic trees and information objects (fragments of RDF-graph). All the rules run "at the same time". This means that in the process of information extraction any rule, for which the input data is sufficient, may be applied at any moment.

The consistency of the extracted information is a built-in feature of the system. It is secured, firstly, by the extraction rules syntax and, secondly, by validation procedures that prevent generation of ontologically inconsistent data.

In addition to RDF graph, extraction mechanism generates annotations, i.e. the information that links extracted data to the respective parts of the original text. In this article we focus on the logical structure of the information extraction mechanism while some details about analysis algorithm may be found in (Starostin et al., 2014).

3.2.1. The rules of interpretation of syntactic-semantic structures

The rules of interpretation of syntactic-semantic structures (the interpretation rules) allow us to find subtrees inside syntactic-semantic trees that meet specific requirements. Each of these rules is a production, in the left part of which a pattern of a tree fragment is defined. In the right part of the production the statements about the information objects, induced by the left part, are grouped. An example of a rule, finding a mention of a person in a syntactic-semantic tree, looks as follows:

```
//Илья Петров пришел
this "PERSON_BY_FIRSTNAME"
[
    surname "PERSON_BY_LASTNAME"
]
=>
Person P(this),
anchor(P, surname, Coreferential),
annotation(P, this.core, surname.core),
P.firstname == Norm(this.core),
P.surname == Norm(surname.core);
```

Figure 1

The most important element of the syntax of syntactic-semantic structures interpretation rules is the construction anchor(...). Thanks to this construction the developer has the possibility to associate information objects with the constituents in a syntactic-semantic tree, which, consequently, allows to state the reference conditions (or object conditions) in the left parts of production rules.

The reference conditions allow to denote the constituents, with which other information objects (by means of other rules) have already been associated. In the following figure an example of a rule with reference condition is given. The rule states that a noun with the meaning "Generalized occupation" (semantic class "HUMAN") refers to the same person that is syntactically dependent on it.

```
//студент Вася
head "HUMAN"
[
    (Classifier_Name|ClassifiedEntity|Specification): this <%BasicEntity:Person%>
]
=>
anchor(this.o,head, Coreferential);
```

Figure 2

The ability to establish multiple links between the world of objects and the text is as important as the ability to establish links between the words within a text. Thanks to this approach different mechanisms of coreference resolution have been organically integrated in our system.

Thus, for example, the system automatically treats the construction anchor(..) in such a way that an object O is automatically linked to the constituents, which are known (on the parser level) to be coreferent with a specific constituent X, while being linked to X. That is why other rules automatically start to "see" the object O on these constituents.

3.2.2. Identification rules

The rules of information object identification allow us to conclude that two information objects, originally considered to be separate, are in fact one and the same object. In contrast to the interpretation rules, these rules are not based on parse trees. The left part of an identification rule contains patterns, describing two information objects (fragments of an RDF graph) and special conditions applied to both objects (for example, that a specific attribute must have the same value). The right part of the rule is empty as the only thing the rule does is merging two objects in one. In the next figure an example of an identification rule merging two organizations into one is given (the reason of merging is the same value of company_by_name attribute).

```
<% Organization, company_by_name ~= null%>
<% Organization, company_by_name ~= null%>

intersects( company_by_name )
```

Figure 3

It is important to note that the constituents to which the information objects are linked are merged along with the merge of the information objects. After that the other rules get the access to the result of the identification process through the united set of constituents, and that, consequently, can lead to the extraction of new information.

We have to leave out the detailed description of anaphoric rules due to limitations on the size of the paper. For information on anaphora and coreference resolution in our system see (Bogdanov et al., 2014).

3.3. Fact Extraction

3.3.1. General Idea

Given the limits of this paper, we chose not to describe the entity extraction module of our system in details. For further information on that please refer to (Starostin et al., 2014). From this point on we will put our focus entirely on fact extraction.

The task of extraction of facts from texts differs considerably from that of named entity recognition. First of all, facts in texts are often expressed non-locally—the information making up a fact (filling specific attributes) is often contained within several sentences. It is possible because natural language provides a variety of instruments to denote coreference—from a simple repeated reference to an object by name to complex types of anaphora. Different combinations of all these instruments can be used to express facts.

It is obvious that a system that solves fact extraction problem efficiently must include coreference resolution mechanisms (Bogdanov et al., 2014). Moreover, the integration of different mechanisms within a single system is a considerable problem. It is well-known that simple architectures with sequential launch of modules (Karasev et al., 2004), responsible for processing of different phenomena of natural language, are confronted with the situation when the information to be generated at late stages is required at the early stages.

The process of fact extraction corresponds to the one described in (Ahn, 2006). There are certain differences however due to the fact that we have a trustworthy semantic parser available.

We define the following stages of fact extraction.

1. Fact occurrence identification. We look for the core of a fact occurrence (in most cases it is a predicate) and create an individual of the fact concept there.
2. Identification of attribute values around the fact core. As our parser provides us with the semantic roles of the words it is unnecessary to differentiate between identification of the attribute values and definition of the attribute type.
   We run our named entity recognition module before launching the fact extraction module. Let us note that the availability of the semantic roles allows us to assign attribute values other than the ontology predefined named entities.
3. Entity and fact coreference resolution.
4. Validation of facts. We filter out the facts that do not have certain attributes filled.

3.4. Specific facts

3.4.1. Occupation

Occupation is the fact that defines personal employment. The key objectives were to identify the employee (most often a person) and the employer (usually an organization or a LocOrg). Both slots were declared mandatory, so should any of them have been missing, the fact was to be discarded. In addition there were two optional properties named position and phase (initiation/termination).

Consider a simplified example of an extraction rule with a single production:

```
    this "TO_WORK" // check if the tree node belongs to "TO_WORK" semantic class, which
combines all sorts of work predicates
        [Agent: person <%BasicEntity:Person%>] // check if "this" has a child with Agent
semantic slot and a Person entity attached to it;
        [(Object|Locative): org <%Org:Organization%>] // check if "this" has a child with
Object or Locative semantic slot and an Organization entity attached to it;
    =>
    BasicFact:Occupation Occ(this); // create Occupation fact
```

Figure 4

It is important to point out that the general model of Occupation in our system is somewhat different from that proposed by the FactRuEval organizers. For instance, we may extract Occupation with no employer, e.g. 'president Barack Obama'. Our system is also capable of extracting implicit positions, for example, in 'Microsoft is led by Bill Gates' we will infer that Bill Gates' position in Microsoft is '(the) head'. Such extra-capabilities were disabled to comply with the competition rules.

3.4.2. Ownership

Ownership generally defines a fact of possession of something (property) by someone (owner). To extract the fact we rely heavily on the 'Possessor' semantic slot ([X] owns Y, Y belongs [to X], [X] bought Y, Y has been acquired [by X], Y was sold [to X], the purchase of Y [by X] etc.). We also make use of semantic classes with 'possessive' meanings and of other elements of our extensive language model.

However, the rules of the competition limited the fact to cases where the property is an Organization and the owner is either a Person or an Organization. To comply with the rules we had to impose certain constraints on the existing rules for Ownership extraction. Some additions were also made for situations where a person was named a (co-)founder of an organization, although the actual *possession* of a company is disputable in this case. Another case that demanded specific adjustments was the ownership of shares.

3.4.3. Deal

Unlike Occupation and Ownership, a Business Deal was not part of the standard set of facts that our system extracts. It turned out, however, that we can cover many cases by simply generalizing several existing facts, Purchase and Transfer above all. This saved us from having to build the extraction library entirely from scratch. Such types of deal as 'loan' and 'investment', on the other hand, required building entirely new rules.

The instances of Purchase fact were converted to Deal if there were Persons or organizations among buyers, sellers and objects of sale. Transfer was specified to Deal only in cases where the object of transfer was a relevant one (e.g. a sum of money). We also considered an organization to be a participant of the deal if its representative person (e.g. an employee or an owner) was already identified as a participant.

The type of the deal was most often determined according to the semantic class of the root node (usually a predicate like 'to buy', 'to invest', 'to be indebted' etc). Sometimes, however, we had to go deeper and examine object under the predicate ('strike [a bargain]', 'sign [a contract]'). We would also like to point out that the test collection contained certain arguable examples of deal types that were not represented in the training set—e.g. economic sanctions, embargoes, fines etc.

4. Dataset

The competition data set was prepared by Opencorpora.org (Bocharov et al.). The document collection consisted of news and analytical articles provided by Wikinews and Chaskor news sites. The markup for track 1 was crowdsourced to the web, tracks 2 and 3 were prepared by FactRuEval organizing committee. The training and the test sets contained 122 and 135 texts respectively.

5. Results

5.1. Entities

The results provided by the organizers show that our system tends to do its best when it comes to Person extraction. On Track 1 we achieved 93% F-measure for Persons and 78,4% for Organizations (Table 1).

Table 1. Track 1 results (test set)

Type	P	R	F1	TP1	TP2	In Std.	In Test.
Per	0.9450	0.9155	0.9300	1,233.18	1,233.18	1,347	1,305
Loc	0.5168	0.8596	0.6455	515.76	515.76	600	998
Org	0.8162	0.7551	0.7844	1,160.57	1,160.57	1,537	1,422
locorg	0.8864	0.3122	0.4618	214.50	214.50	687	242
overall	**0.7875**	**0.7490**	**0.7678**	3,124.01	3,124.01	4,171	3,967

Table 2. Track 1 results without Loc/LocOrg division (test set)

Type	P	R	F1	TP1	TP2	In Std.	In Test.
per	0.9450	0.9155	0.9300	1,233.18	1,233.18	1,347	1,305
loc	0.9261	0.8698	0.8971	1,116.83	1,116.83	1,284	1,206
org	0.8175	0.7564	0.7858	1,162.56	1,162.56	1,537	1,422
overall	**0.8931**	**0.8427**	**0.8672**	3,512.57	3,512.57	4,168	3,933

Locations clearly fell victim to the Location/LocOrg split, which was quite hard to formalize and implement to begin with. As can be seen from the table, the relatively moderate quality of the Location extraction is mainly due to a lot of false positives, and this drop in precision corresponds to LocOrg's drop in recall. To put it simply, the organizers' understanding of when a Location becomes a LocOrg was apparently much broader than ours. There was a mode of comparison without Loc/LocOrg division (Table 2), where this split was removed. Our system, actually, ranked 1 in this mode of comparison.

Table 3 shows the results of Track 2. Note that in this case all results fall into the same precision-better-than-recall pattern typical of rule-based systems.

Table 3. Track 2 results (test set)

Type	P	R	F1	TP1	TP2	In Std.	In Test.
Per	0.8817	0.8592	0.8703	538.73	538.73	627	611
Loc	0.8430	0.7942	0.8179	494.00	494.00	622	586
Org	0.6823	0.6763	0.6793	547.17	547.17	809	802
overall	**0.7903**	**0.7677**	**0.7789**	1,579.90	1,579.90	2,058	1,999

We also noted that our system tends to demonstrate marginal or no decline in quality when we shift to the test set from the training one. For Track 1 the overall F-measure actually even went up on the test set (from 75.7% to 76.8%), and for Track 2 the drop was not dramatic (from 83.2% to 77.9%).

5.2. Facts

Facts have a more sophisticated and variable structure than entities, they are much more syntax-dependent and tend to 'spread' across large sections of text (at times much larger than a single sentence). This makes it hard to detect every participant of a single fact and account for all possible patterns and paraphrases. The results of the third track show that while we were able to achieve good precision for most facts, there is much room for improvement when it comes to recall.

Table 3. Track 3 results (test set)

TAG	P	R	F1	TP1	TP2	In Std.	In Test.
ownership	0.5379	0.1709	0.2594	24.10	26.90	141	50
occupation	0.8058	0.5679	0.6662	190.80	195.81	336	243
meeting	0.8690	0.1352	0.2340	6.08	6.08	45	7
Deal	0.6777	0.1932	0.3007	19.71	27.11	102	40
overall	**0.7526**	**0.3857**	**0.5100**	240.69	255.89	624	340

To our knowledge, these were the best results for Track 3. However, it should be noted that only two systems took part in track 3.

6. Conclusion

FactRuEval-2016 allowed us to evaluate the performance of our Information Extraction system in a competitive environment, and we are grateful to the organizing committee for this opportunity.

The results for entity extraction show that our system has a slight bias for precision over recall, which is typical for rule/pattern-based approaches. Overall, we were able to achieve good quality (especially with Persons) comparable with the results of the best machine learning systems. Moreover, our approach proved to be quite stable, showing little or no decline in F-measure

Fact extraction results turned out to be even more precision-oriented: the system returns few incorrect fact, but also misses a fair share of correct ones. However, the use of syntactic-semantic trees helped us create concise rules that covered a big subset of possible patterns for the four facts chosen by the organizing committee. And as facts depend heavily on syntactic structures, we faced very little competition from the machine learning based systems.

Analysis of mistakes suggests that further work should be concentrated on sophisticated coreference resolution (Vladimir Putin—president—leader—head). Some facts from the test set clearly required the system to make use of document-level information and pragmatics of the text, like in cases when a president or prime minister strikes a deal, and the country(s) he implicitly represents is listed as a participant.

References

1. *Ahn D.* (2006), The stages of event extraction, Proceedings of the Workshop on Annotating and Reasoning about Time and Events, Association for Computational Linguistics, pp. 1–8.
2. *Anisimovich K. V., Druzhkin K. Ju., Minlos F. R., Petrova M. A., Selegey V. P., Zuev K. A.* (2012), Syntactic and semantic parser based on ABBYY Compreno linguistic technologies, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog” [Komp’iuternaia

Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"], Bekasovo, pp. 90–103.
3. *Bocharov V., Bichineva S., Granovsky D., Ostapuk N., Stepanova M.* (2011), Quality assurance tools in the OpenCorpora project, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"], Bekasovo, Vol. 10.
4. *Bogdanov A. V., Dzhumaev S. S., Skorinkin D. A., Starostin A. S.* (2014), Anaphora Analysis based on ABBYY Compreno Linguistic Technologies, Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2014), pp. 659–667.
5. *Bunescu R., Mooney R. J.* (2004), Collective information extraction with relational Markov networks, Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, p. 438.
6. *Califf M. E., Mooney R. J.* (2003), Bottom-up relational learning of pattern matching rules for information extraction, The Journal of Machine Learning Research, Vol. 4, pp. 177–210.
7. *Chieu H. L., Ng H. T., Lee Y. K.* (2003), Closing the gap: Learning-based information extraction rivaling knowledge-engineering methods, Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Vol. 1, pp. 216–223.
8. *Freitag D.* (1998), Toward general-purpose learning for information extraction, Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Association for Computational Linguistics, Vol. 1, pp. 404–408.
9. *Freitag D., McCallum A.* (2000), Information extraction with HMM structures learned by stochastic optimization, AAAI/IAAI, pp. 584–589.
10. *Gareev R., Tkachenko M., Solovyev V., Simanovsky A., Ivanov V.* (2013), Introducing baselines for Russian named entity recognition, Computational Linguistics and Intelligent Text Processing, Springer Berlin Heidelberg, pp. 329–342.
11. *Grishman R., Sundheim B.* (1996), Message Understanding Conference-6: A Brief History, COLING, Vol. 96, pp. 466–471.
12. *Hogenboom F., Frasincar F., Kaymak U., De Jong F.* (2011), An overview of event extraction from text, Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) at Tenth International Semantic Web Conference (ISWC 2011), Vol. 779, pp. 48–57.
13. *Karasev V., Khoroshevsky V., Shafirin A.* (2004), New Flexible KRL JAPE+: Development & Implementation, Knowledge-Based Software Engineering: Proceedings of the Sixth Joint Conference on Knowledge-Based Software Engineering, IOS Press, Vol. 108, p. 217.
14. *Mel'čuk I. A.* (1973), Towards a Linguistic 'Meaning⇔Text' Model, Trends in Soviet theoretical linguistics, Springer Netherlands, pp. 33–57.
15. *Nadeau D., Sekine S.* (2007), A survey of named entity recognition and classification, Lingvisticae Investigationes, Vol. 30, pp. 3–26.

16. *Patwardhan S., Riloff E.* (2009), A unified model of phrasal and sentential evidence for information extraction, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Vol. 1, pp. 151–160.
17. *Riloff E.* (1996), Automatically generating extraction patterns from untagged text, Proceedings of the national conference on artificial intelligence, pp. 1044–1049.
18. *Shelmanov A., Smirnov I., Vishneva R.* (2015), Information Extraction from Clinical Texts in Russian, Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"].
19. *Soderland S.* (1999), Learning information extraction rules for semi-structured and free text, Machine learning, Vol. 34, pp. 233–272.
20. *Soderland S., Fisher D., Aseltine J., Lehnert W.* (1995), CRYSTAL: Inducing a conceptual dictionary, arXiv preprint cmp-lg/9505020.
21. *Solovyev V., Ivanov V., Gareev R., Serebryakov S., Vassilieva N.* (2012), Methodology for Building Extraction Templates for Russian Language in Knowledge-Based IE Systems, HP Laboratories Technical report, HPL-2012–211.
22. *Starostin A. S., Smurov I. M., Stepanova M. E.* (2014), A Production System for Information Extraction Based on complete syntactic-semantic analysis, Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2014), pp. 89–101.
23. *Tjong Kim Sang E. F., De Meulder F.* (2003), Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, Association for Computational Linguistics, Vol. 4, pp. 142–147.
24. *W3C* (2004), OWL Web Ontology Language Overview, available at: http://www.w3.org/TR/2004/REC-owl-features-20040210
25. *W3C* (2006), Defining N-ary Relations on the Semantic Web, available at: http://www.w3.org/TR/swbp-n-aryRelations
26. *Yakushiji A., Miyao Y., Ohta T., Tateisi Y., Tsujii J. I.* (2006), Automatic construction of predicate-argument structure patterns for biomedical information extraction, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 284–292.
27. *Yangarber R., Grishman R., Tapanainen P., Huttunen S.* (2000), Automatic acquisition of domain knowledge for information extraction, Proceedings of the 18th conference on Computational linguistics, Association for Computational Linguistics, Vol. 2, pp. 940–946.
28. *Zuev K. A., Indenbom M. E., Judina M. V.* (2013), Statistical machine translation with linguistic language model, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"], Bekasovo, Vol. 2, pp. 164–172.

КОНТРОЛЬ БЕССОЮЗНОГО ЦЕЛЕВОГО ИНФИНИТИВА ПРИ ГЛАГОЛАХ КАУЗАЦИИ ДВИЖЕНИЯ В РУССКОМ ЯЗЫКЕ: ДАННЫЕ НКРЯ

Стойнова Н. М. (stoynova@yandex.ru)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

В работе на корпусном материале рассматриваются русские целевые конструкции с глаголом каузации движения и бессоюзным инфинитивом. Контролировать бессоюзный инфинитив при глаголе каузации движения типа *вести* может субъект или объект финитного глагола (*повел его купать* vs. *повел его купаться*). Склонность к субъектному контролю коррелирует с большей формальной и семантической спаянностью инфинитивной конструкции, склонность к объектному контролю, наоборот, с меньшей. Этим, в частности, объясняются статистические корреляции, наблюдаемые на материале корпуса: различия в соотношении употреблений с субъектным / объектным контролем при разном порядке слов и при разных пространственных типах глагола каузации движения. Выбор между субъектным / объектным контролем определяется также семантикой объекта (его положением на шкале одушевленности) и семантикой целевой ситуации.

Ключевые слова: глаголы движения, целевые конструкции, инфинитивные конструкции, движение с целью

CONTROL IN PURPOSE CONSTRUCTIONS WITH CAUSATION-OF-MOTION VERBS IN RUSSIAN: EVIDENCE FROM RUSSIAN NATIONAL CORPUS

Stoynova N. M. (stoynova@yandex.ru)

V. V. Vinogradov Russian Language Institute, Moscow, Russia

The paper deals with motion-cum-purpose constructions in Russian. The constructions “causation-of-motion verb + infinitive” (*vesti kogo-to delatj čto-to* — lit. ‘to lead somebody to do something’) are observed on the data of Russian National Corpus. The problem of infinitive control is in focus. The Ø-subject of infinitive can be coreferential with the subject or with the

object of the finite verb in such a construction: cf. *on vedet jejo ubivatj* lit. 'he$_i$ is leading her$_j$ Ø$_i$ to kill' vs. *on vedet jejo umiratj* lit. 'he$_i$ is leading her$_j$ Ø$_j$ to die'. The frequency of uses with subject-control correlates with the degree of formal and semantic cohesion within the purpose construction. One of the parameters is word order (which reflects the degree of syntactic cohesion). The second one is the spatial type of causation-of-motion verb: neutral unprefixed (*vesti*) / lative prefixed (*privesti*) / elative prefixed (*otvesti*). It reflects somehow the degree of semantic cohesion (neutral > lative > elative). The choice of controller is conditioned also by the referential type of the object (person / non-person / inanimate) and by the semantics of purpose event.

Key words: motion verbs, purpose clauses, infinitive constructions, motion-cum-purpose

1. Глаголы каузации движения и проблема контроля инфинитива

В работе на материале Национального корпуса русского языка (ruscorpora.ru, далее НКРЯ) будут рассмотрены некоторые случаи вариативности в контроле бессоюзного целевого инфинитива при глаголах каузации движения. Имеются в виду примеры вроде (1) и (2).

(1) *После ужина всех девочек* ***повели мыться***. (2011)[1]

(2) *Он был в переделке, сейчас его* ***повели мыть****, отпаривать, смазывать…* (1998)

В (1) невыраженный субъект инфинитива кореферентен объекту глагола каузации движения (*мыться* будут *девочки*) — объектный контроль инфинитива (далее О-контроль); тогда как в (2) — субъекту глагола каузации движения (*мыть* будут те, кто *повели*) — субъектный контроль (далее А-контроль). В отличие от бессоюзного инфинитива, инфинитив, вводимый союзом *чтобы*, при тех же глаголах контролируется почти исключительно субъектом (об отдельных исключениях см. [Пекелис 2002: 58–59]) и далее не рассматривается.

Бессоюзный инфинитив со значением цели возможен при небольшой, но достаточно разнородной группе глаголов (ср., например, *пойти*, *дать*, *взять*, *оставить*, *повесить*, *отвести*, *отнести* и др.), подробный список и обсуждение см. в [Журавлева 1999]; [Градинарова 2006]. Правила его контроля достаточно сложны, не универсальны и для разных глаголов требуют отдельного подробного рассмотрения. Контролировать инфинитив в зависимости от разных условий могут субъект финитного глагола (*пошел мыться*, *повел мыть*), объект — для глаголов с валентностью на объект (*повел меня мыться*) и адресат — для глаголов с адресатной валентностью (*дал мне посмотреть*).

[1] Все примеры ниже взяты из НКРЯ.

В общем виде такого рода правила описаны в [Пекелис 2002: 44–55]. В работе глаголы, допускающие бессоюзный целевой инфинитив, разделены на четыре семантических класса, для каждого из которых формулируются отдельные правила. При этом оговаривается возможность отнесения одного и того же глагола в разных употреблениях к разным классам, и в пример глагола, относящегося сразу к двум группам (т.е., иными словами, не подчиняющегося единому правилу), приводится как раз один из глаголов каузации движения, подробно рассматриваемых ниже.

Наблюдаемая вариативность в контроле инфинитива при глаголах типа *вести* интересна в более широком контексте описания целевых конструкций при глаголах движения (и каузации движения как частного случая). Известно, что глаголы движения представляют собой с точки зрения целевых конструкций особую группу. Прагматически целевая ситуация более тесно связана с ситуацией движения, чем с другими типами ситуаций. И в языках мира это может находить отражение в т.ч. на поверхностном уровне. Ср. особые формы или синтаксические конструкции для оформления цели при глаголе движения, грамматикализация глагола движения в целевой конструкции во вспомогательный или полувспомогательный, деривационные аффиксы со значением ‘пойти сделать V’, грамматикализация целевой конструкции с глаголом движения в видо-временную (ср. англ. *to be going to*), подробнее см., например, [Schmidtke-Bode 2009: 93ff.].

Для русского языка свидетельства особого статуса глаголов движения в целевой конструкции не настолько наглядны, однако признаки большей формальной и семантической спаянности целевой конструкции с глаголами движения также обнаруживаются. В частности, показателен сам факт возможности оформления цели при глаголах движения (наряду с очень немногими другими) бессоюзным инфинитивом, а не обычной союзной клаузой со *чтобы*. Про такие конструкции встает, во-первых, вопрос о том, проявляют ли они свойства обычной биклаузальной конструкции или частично сходны с моноклаузальными. Во-вторых, встает вопрос о степени интеграции инфинитива в структуру глагола движения. Если для целевой клаузы при обычном глаголе очевиден ее сирконстантный статус, то для бессоюзного инфинитива при глаголе движения правомерен вопрос об актантном статусе или статусе, промежуточном между сирконстантным и актантным, см. [Гусев 2004].

С точки зрения степени формальной и семантической спаянности целевой конструкции можно рассматривать и вопрос о контроле инфинитива (аналогичным образом он рассматривается, например, в [Aissen 1984] для подобной конструкции с глаголами движения в цоциль). Если при одновалентном глаголе движения типа *идти* контролировать инфинитив может только его единственный актант (субъект), и тем самым вопроса не возникает, то при двухвалентном глаголе каузации движения типа *вести* (далее КД), как было показано выше, кандидатов два — субъект и объект каузативного глагола, причем жесткого синтаксического правила выбора между ними нет. Можно предположить, что субъектный контроль — при котором нет рассогласования между субъектом финитного глагола и инфинитива — будет коррелировать с признаками большей спаянности конструкции, а объектный — с признаками меньшей спаянности.

1.1. Контроль инфинитива при глаголах каузации движения: некоторые дополнения

Выбор контролера инфинитива при рассматриваемых нами глаголах КД, как сказано выше, происходит между объектом и субъектом глагола КД. Это, однако, нуждается в некоторых оговорках.

Иногда совместными исполнителями действия, названного инфинитивом, являются de facto и объект, и субъект глагола КД, как в (3), или объект и участник с ролью адресата/конечного пункта движения, как в (4):

(3) …***Ее*** *между тем* ***увел вальсировать*** *Чиаурели. — вальсировать* будут и героиня (*она*), и *Чиаурели* (1980–1998)

(4) …***меня привели знакомиться*** *к одной очень талантливой художнице…* (2012) — *знакомиться* будут *говорящий* и *художница*

Все такие примеры в силу неоднозначности их интерпретации условно отнесены к случаям О-контроля.

Некоторые случаи, интерпретируемые ниже как случаи А-контроля, на самом деле могут предполагать реализацию ситуации третьим лицом, см. (5)[2].

(5) *И буря негодования вспыхнула во мне:* ***ведут судить!*** (1906) — *судить* будут, вероятнее всего, не те, кто *ведут*

При О-контроле глагол в инфинитиве может быть непереходным, а может быть переходным и иметь выраженный прямой объект (*выводили арестантов убирать площадь*). При А-контроле глагол в инфинитиве переходный, объект при нем не выражен и совпадает с объектом глагола КД. Это уже само по себе говорит о том, что конструкция с А-контролем проявляет больше признаков формальной спаянности, чем конструкция с О-контролем.

Особым образом ведет себя глагол *показ*(*ыв*)*ать*. Для него возможно два типа употреблений с А-контролем. Первое — стандартное, при котором объект глагола КД тождествен невыраженному объекту глагола *показать*, см. (6); второе — значительно более частотное (62 из 66 употр.) — такое, при котором объекту глагола КД соответствует не объект глагола *показать*, а его невыраженный адресат, см. (7):

(6) *…но она крепилась и даже* ***привела меня показать*** *московским родственникам, сотрудникам КГБ.* (1997) — *показать меня*

---

[2] Чуть проще случай «расширенного» А-контроля и О-контроля глаголов обслуживания типа *повела сына стричь / стричься в парикмахерскую* (упоминаемый в т.ч. в [Пекелис 2002]), в котором ситуация реализуется также не собственно субъектом / объектом, но которому можно поставить в соответствие такие же расширенные финитные употребления *стрижет сына в парикмахерской / стрижется в парикмахерской*.

(7) — ***Гостей**, прилетающих в Мирный, в первый же день **ведут показать** пингвинов.* (1964) — *показать гостям*

Различаются правила контроля для ситуаций с неодушевленным vs. одушевленным объектом при глаголе КД. При неодушевленных объектах выбирается почти исключительно О-контроль: примеры типа (8) единичны (хотя и грамматичны).

(8) *Потом охотники поднимают байдару на плечи и **несут сушиться**…* (1989)

Зато для ситуаций с неодушевленным объектом возможен, помимо А-контроля, контроль со стороны адресата/конечного пункта глагола КД:

(9) *…одна женщина **принесла мне почитать** толстую книгу «Летопись Серафимо-Дивеевского монастыря».* (2004)

Для ситуаций с одушевленным объектом контроль адресата не кажется невозможным, однако в выборке из НКРЯ очевидных примеров такого рода не встретилось. Ср. (10), который можно трактовать и как случай «расширенного» А-контроля:

(10) *Одно звание — кузнецы, а **приведут лошадь ковать**, угля нет, купить не на что.* ((1899) — *ковать* будут *кузнецы* (к которым привели лошадь) или те, кто *привели*, — *у кузнецов*

Ниже подробно обсуждается только — менее предсказуемый — выбор контролера в ситуациях с одушевленным объектом.

1.2. Материал исследования

Исследование проведено на сплошной выборке примеров из Основного корпуса НКРЯ на бессоюзный инфинитив при глаголе каузации движения, всего 672 примера[3]. Поиск с последующей ручной фильтрацией велся по запросам типа: а) *вести*+0…0+inf (контактная позиция глагола каузации движения с инфинитивом); б) *вести*+0…0+acc+0…0+inf (объект при глаголе движения в интерпозиции)[4].

---

[3] На равных основаниях, таким образом, рассматривались все употребления, начиная с XVIII в. Достаточно большой временной разброс следует принять во внимание при интерпретации обсуждаемых результатов. Серьезных диахронических сдвигов в этом фрагменте грамматики на первый взгляд в целом не наблюдается; это, однако, требует дополнительной проверки.

[4] Таким образом, в выборку не вошли примеры типа (i) с бóльшим линейным расстоянием между финитным глаголом и инфинитивом, а также типа (ii) с препозицией инфинитива: (i) *Я тут же подбежал к нему, помог сойти с плотика и **повел** к печи **показать**, подойдет ли размер основания печи.* (2010) — (ii) *Как его **казнить повели**, и мать в белом покрывале на балконе стояла.* (1970).

Из всех глаголов каузации движения были рассмотрены только глаголы с корнем *-вес(ти)*: *вести*, *повести*, *отвести*, *привести*, *увести*, *вывести*[5]. Такое ограничение материала обусловлено тем, что *вести* и его дериваты предполагают одушевленный объект, тогда как другие глаголы каузации движения (*нести*, *везти*, *тащить*) — преимущественно или в том числе и неодушевленный, для конструкции с которым правила контроля более просты, см. выше.

2. Порядок слов в инфинитивной конструкции с глаголом каузации движения

Количественное соотношение употреблений с А-контролем и с О-контролем разное при разном порядке слов. Рассмотрены три типа порядка слов:

а) Объект после глагола КД и перед инфинитивом (V+acc+inf):

(11) *в огромной квартире отчима Сергея, куда тот* ***повел Августа обедать***. (2001)

(12) *Я* ***повела их кормить*** *в столовую…* (1984)

б) Объект в препозиции к глаголу КД (acc+V+inf):

(13) *После свидания с родными* ***девочек повели обедать***. (1894)

(14) *С этого времени жить стало лучше.* ***Его повели купаться***. (1918–1919)

(15) *Вот мы, говорит,* ***жеребцов заводских поведем купать****, и тебя с ними!* (1867)[6]

в) Объект в постпозиции к инфинитиву (V+inf+acc):

(16) *Дрессировщик из цирка* ***ведет купаться медведей***. (1985)

(17) *…Павел отправился купаться и* ***повел купать*** *на реку* ***лошадей***. (1945–1955)

Конструкция а) с интерпозицией объекта — линейно расщепленная, конструкция в) — максимально спаянная. Для нее, собственно, неочевидно, трактовать ли объект как относящийся к глаголу КД, к инфинитиву или к тому и другому сразу — как к единому глагольному комплексу. Конструкция в)

---

[5] Для других дериватов *вести* (как *ввести*, *завести*) в НКРЯ вообще нет примеров с инфинитивом или они единичны.

[6] К этой же группе условно отнесены употребления типа *взял ее и повел купать* и единичные без выраженного в пределах предложения объекта.

отличается от а) и б) также большей коммуникативной спаянностью: объект в ней входит в рему — как правило, вместе с глаголом КД и инфинитивом.

Вполне естественно, что доля употреблений с А-контролем оказывается самой большой для конструкции в) и самой малой для конструкции а), см. Таблицу 1.

Таблица 1. Порядок слов в инфинитивной конструкции с глаголом КД[7]

	О-контроль	**А-контроль**	**% с А-контролем**
V+acc+inf	257	72	22%
acc+V+inf	187	114	38%
V+inf+acc	14	28	67%
всего	**458**	**214**	**32%**

3. Пространственный тип глагола каузации движения

Доля употреблений с А-контролем и с О-контролем зависит от конкретного глагола КД, для всех глаголов (предполагающих одушевленный объект) при этом доля употреблений с О-контролем превышает долю употреблений с А-контролем, см. Таблицу 2:

Таблица 2. Разные глаголы с корнем *-вес(ти)*

	О-контроль	**А-контроль**	**% с А-контролем**
вести	64	55	46%
повести	219	129	37%
привести	62	17	22%
увести	36	5	12%
вывести	61	7	10%
отвести	16	1	6%

Картину, представленную в Таблице 2, можно увязать с пространственной семантикой глагола КД. Ср. классификацию глаголов движения в [Падучева 2004: 373–384]: 1) бесприставочные (*идти*) и примыкающие к ним глаголы с непространственной приставкой (*пойти*) — нейтральные; 2) глаголы с приставкой, акцентирующей конечный пункт: *в, вз, за, до, на, под, при* (*войти, прийти*) — лативные; 3) глаголы с приставкой, акцентирующей исходный пункт: *вы, от, с, у…* (*отвести, увести, вывести*) — элативные. По частотности уопотреблений с А-контролем глаголы упорядочены следующим образом: нейтральные > лативные > элативные, см. Таблицу 3.

[7] Все различия статистически значимы, критерий $\chi 2$, $p < 0{,}01$.

Таблица 3. Разные пространственные типы глаголов[8]

	О-контроль	**А-контроль**	**% с А-контролем**
нейтральные (*вести*, *повести*)	283	184	**39%**
лативный (*привести*)	62	17	**22%**
элативные (*увести, отвести*, *вывести*)	113	13	**10%**

Для объяснения этой асимметрии важны следующие моменты. 1) Как отмечалось в [Гусев 2004] и [Градинарова 2006], бессоюзная целевая конструкция чувствительна к пространственной смежности ситуации движения и ситуации-цели: цель, выражаемая инфинитивом, должна быть реализована в конечном пункте движения (ср. **он ушел избежать неприятной встречи*). Подобное ограничение, хотя и менее жесткое, касается, видимо, и временной смежности: наличие временной дистанции между ситуациями движения и цели как минимум не должно акцентироваться, а в идеале должно отсутствовать (**он только что приехал завтра дать концерт*). 2) Соответственно, участниками целевой ситуации, описываемой бессоюзной инфинитивной конструкцией, в обычном случае должны стать только те референты, которые на момент окончания (результирующей фазы) ситуации движения находятся в конечном пункте движения. 3) Из двух участников ситуации каузации движения — субъекта (каузатора) и объекта (каузируемого) — объект оказывается участником целевой ситуации в любом случае, а субъект может концептуализоваться как участник целевой ситуации (при А-контроле) или оставаться за кадром (при О-контроле). Таким образом, наибольшей склонности к А-контролю, концептуализующему обоих участников ситуации движения как участников целевой ситуации, мы можем ожидать в случае максимального соблюдения пространственно-временного единства обоих участников.

В этом месте и обнаруживается различие между нейтральными, лативными и элативными глаголами КД. Элативный одноместный глагол движения типа *отойти* выводит участника из точки наблюдения (точка наблюдения — исходный пункт движения), см. [Падучева 2004: 379ff.]. Для глагола каузации движения из точки наблюдения выводится объект, но не / необязательно субъект — они, следовательно, оказываются пространственно разделены. Особенно этот эффект очевиден для глагола *отвести*: *отвел ≈ увел и вернулся*. Глаголы *увести* и *вывести* не содержат импликации о том, что субъект возвращается в точку наблюдения, но и не содержат информации о том, что он остается в той же точке, что и объект. Для лативного *привести* точка наблюдения — конечный пункт движения (где в момент наблюдения оказываются вместе субъект и объект), для *вести*, *повести* она не фиксирована, и эффекта

[8] Все различия статистически значимы, критерий χ^2, $p < 0{,}05$.

пространственного разделения субъекта и объекта каузации движения не возникает. Разница между нейтральными глаголами и лативными связана не с пространственной, а с временной разделенностью субъекта и объекта. Для лативного *привести* (а также элативных) в фокусе результирующая фаза ситуации движения, когда ситуация движения (с одним набором участников) уже завершилась, а целевая ситуация (с, возможно, другим набором участников) еще не началась. Для нейтральных глаголов *вести* и *повести* в фокусе начальная или срединная фаза ситуации, а результирующая фаза ситуации движения, так же, как и вся целевая ситуация, отнесены к плану будущего, т.е. между ними (и их потенциальным набором участников) сохраняется временное единство. Все это, как кажется, и объясняет наблюдаемую асимметрию в склонности к А-/О-контролю, см. суммирующую Таблицу 4.

Таблица 4. Единство участников ситуации движения и целевой ситуации: разные пространственные типы глаголов КД

	единство А и О в пространстве	**единство А и О во времени**
нейтральные	+	+
лативные (*привести*)	+	–
элативные	–	–

Указанные пространственные типы глаголов движения соотносятся также с их аргументной структурой — обязательностью / факультативностью участника конечный пункт, однако объяснение через аргументную структуру глагола КД и актантный / сирконстантный статус инфинитива кажется менее убедительным. Косвенное тому подтверждение — отсутствие корреляции между А-/О-контролем и наличием в предложении предложной группы с ролью конечного пункта (*привести обедать / привести обедать в ресторан*), см. Таблицу 5.

Таблица 5. Глагол *повести*: с выраженным конечным пунктом и без выраженного конечного пункта[9]

	О-контроль	**А-контроль**	**% с А-контролем**
+goal	29	17	37%
–goal	187	115	38%

[9] Различие статистически незначимо, χ^2, p = 0,8837.

4. Семантика инфинитива и семантический тип участников ситуации

Следующий параметр касается семантики ситуации-цели, вводимой инфинитивом, и типом ее участников, в первую очередь того, который является объектом ситуации каузации движения. Уже упоминалось, что особым образом ведут себя не рассмотренные тут конструкции с неодушевленным объектом: для них О-контроль практически невозможен. Понижено количество употреблений с О-контролем и для конструкций с одушевленным объектом нелицом (животные): ср. разницу в Таблице 6.

Таблица 6. Объект глагола КД: лицо / не-лицо одуш.[10]

	О-контроль	**А-контроль**	**% с А-контролем**
не-лицо	23	35	60%
лицо	435	179	29%
всего	**458**	**214**	**32%**

Что касается семантики ситуации, вводимой инфинитивом, то круг глаголов, наблюдаемых при А-контроле и при О-контроле, очень различается, и лишь для немногих употреблений существует абсолютный или хотя бы приблизительный эквивалент с разницей только в синтаксическом контроле. Даже пара типа (18)–(19) с глаголами *знакомить — знакомиться* не синонимична: в (18) предполагается, что субъект глагола КД непосредственно участвует в процессе знакомства, в (19) необязательно:

(18) *Она схватила его за руку и повела* ***знакомить*** *со своими подружками…* (1977)

(19) *Потом Ляля пошла наверх по скрипучей лестнице в светелку — старушка* ***повела знакомиться*** *к больной девочке.* (1971)

При А-контроле существенную часть употреблений (32%, 47 употр.) составляют глаголы со значением кары (*расстреливать*, *казнить*, *вешать* и под.): их существенное свойство — резкая асимметрия между субъектом (каузатором) и объектом (каузируемым) ситуации (ситуации со «сверхпассивным пациенсом»). К ним примыкают другие подобные, напр., *продавать*.

При О-контроле около половины употреблений (49%, 96 употр.) составляют глаголы «совместной деятельности», обозначающие ситуации, обязательно или часто предполагающие двух или более участников (*танцевать*, *обедать <в ресторане>*, *гулять <по городу>*). Один из них в конструкции с глаголом КД представлен как каузатор, а другой/другие как каузируемый, однако на самом деле они оба они участвуют в ситуации, обозначенной инфинитивом, в одной и той же роли: т.е. для них, наоборот, наблюдается симметрия между каузатором и каузируемым.

[10] Различие статистически значимо, χ^2, $p < 0{,}0001$.

5. Разрозненные наблюдения над контролем инфинитива и степенью спаянности конструкции: примеры из НКРЯ

Ниже приводятся отдельные примеры из НКРЯ, также интересные в контексте проблематики контроля инфинитива при глаголах КД и спаянности инфинитивной конструкции.

1) Сочинение инфинитивов с О-контролем и А-контролем при одном глаголе КД:

(20) …*чтобы она уже* ***увела их пить*** <О-контроль> *кофе* ***и кормить*** <А-контроль> *своими собственными бутербродами.* (2009)

2) Рефлексивное местоимение в составе инфинитивного оборота при О-контроле, кореферентное субъекту КД (но не инфинитива):

(21) *Потом* $Ø_j$ *повела меня*$_i$ $Ø_i$ *смотреть своего*$_j$ *сына…* (1910)

3) Двойное выражение объекта / адресата при глаголе *показывать*[11]:

(22) *После обеда пришел барон Криднер, и я же* ***повел его***$_i$ ***показывать ему***$_i$ *город и окрестности.* (1855)

4) Локативный участник при инфинитиве с А-контролем, наследующий модель управления от глагола КД:

(23) *Был назначен вторично, то его приняли весьма сурово, много били по щекам, таскали за волосы и повели топить в Енисей.* (1916) — **топили в Енисей* / ^{ОК}*в Енисее*; *#повели в Енисей*

5) Аномальные с т.зр. коммуникативной структуры и порядка слов примеры с А-контролем (выраженный объект при инфинитиве, а не глаголе КД):

(24) *И все они* ***ведут побивать ее камнями.*** (1887) — ср. ^{ОК}*идут побивать ее камнями*

6. Заключение

Таким образом, на материале НКРЯ обнаруживаются следующие тенденции.

а) Доля употреблений с А-/О-контролем коррелирует с линейной позицией объекта: конструкции, проявляющие большую спаянность на уровне порядка слов, легче допускают А-контроль.

[11] Все такие примеры относятся к XIX в.

б) Доля употреблений с А-/О-контролем коррелирует с пространственным типом глагола КД: больше всего употреблений с А-контролем в конструкциях с нейтральными глаголами *вести, повести*, меньше в конструкциях с лативными глаголами, менее всего — с элативными. Это объясняется степенью семантической спаянности ситуации движения с целевой ситуацией, разной для этих типов, а именно пространственно-временным единством их участников.

в) Доля употреблений с А-/О-контролем коррелирует с семантическим типом целевой ситуации и положением О-участника на шкале одушевленности: легче допускается А-контроль для не-личных участников и для асимметричных ситуаций со «сверхпассивным» пациенсом; значительная часть конструкций с О-контролем вводит, наоборот, целевые симметричные ситуации «совместного действия».

Литература

1. *Градинарова А.* (2006), Русский целевой бессоюзный инфинитив: условия употребления // Болгарская русистика, 3–4, сс. 11–20.
2. *Гусев В. Ю.* (2004), Целевые конструкции при глаголах движения: актанты или сирконстанты? // International symposium on typology of the argument structure and grammatical relations in languages spoken in Europe and North and Central Asia, Kazan State University, May 11–14.
3. *Журавлева О. Н.* (1999), Семантика и фукционирование конструкций, включающих глаголы движения, перемещения, изменения положения в пространстве и инфинитив цели, Дисс. канд. филол. наук, Киров.
4. *Падучева Е. В.* (2004), Динамические модели в семантике лексики. Языки славянских культур, Москва.
5. *Пекелис О. Е.* (2002), Субъект зависимого инфинитива в русском и итальянском языках (проблема контроля). Дипломная работа. М.: РГГУ, Москва.
6. *Aissen J.* (1984), Control and command in Tzotzil purpose clauses // Berkeley Linguistics Society, Vol. 10, pp. 559–71.
7. *Schmidtke-Bode K.* (2009), A Typology of Purpose Clauses, John Benjamins, Amsterdam–Philadelphia.

References

1. *Aissen J.* (1984), Control and command in Tzotzil purpose clauses // Berkeley Linguistics Society, Vol. 10, pp. 559–71.
2. *Gradinarova A.* (2006), Russian purpose infinitive without conjunction: the conditions of use [Russkij celevoj bessojuznyj infinitiv: uslovija upotreblenija], Bolgarskaja rusistika, Vol. 3–4, pp. 11–20.

3. *Gusev V. Ju.* (2004), Purpose constructions with motion verbs: arguments or adjuncts? [Celevyje konstrukcii pri glagolah dviženija: aktanty ili sirkonstanty?] // International symposium on typology of the argument structure and grammatical relations in languages spoken in Europe and North and Central Asia, Kazan, Kazan State University.
4. *Padučeva E. V.* (2004), Dynamic models in semantics of lexicon [Dinamičeskije modeli v semantike leksiki], Jazyki slavjanskih kuljtur, Moscow.
5. *Pekelis O. E.* (2002), Subject of dependent infinitive in Russian and Italian (the problem of control), [Subjekt zavisimogo infinitiva v russkom I italjanskom jazykah (problema kontrolja)], Master thesis, RSUH, Moscow.
6. *Schmidtke-Bode K.* (2009), A Typology of Purpose Clauses, John Benjamins, Amsterdam–Philadelphia.
7. *Zhuravleva O. N.* (1999), Semantics and functions of purpose infinitive constructions with verbs of motion and change of position [Semantika I funkcionirovanije konstrukcij, vključajušjih glagoly dviženija, peremešjenija, izmenenija položenija v prostranstve I infinitiv celi], Diss. kand. filol.nauk, Kirov.

РАСПОЗНАВАНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ: ПОДХОД НА ОСНОВЕ ВИКИ-РЕСУРСОВ

Сысоев А. А. (sysoev@ispras.ru),
Андрианов И. А. (ivan.andrianov@ispras.ru)

Институт системного программирования
РАН, Москва, Россия

Ключевые слова: распознавание именованных сущностей, вики-ресурсы, машинное обучение, векторное представление слов

NAMED ENTITY RECOGNITION IN RUSSIAN: THE POWER OF WIKI-BASED APPROACH

Sysoev A. A. (sysoev@ispras.ru),
Andrianov I. A. (ivan.andrianov@ispras.ru)

Institute for System Programming of RAS, Moscow, Russia

Named entity recognition and classification is an important natural language processing task, aimed at finding words and word sequences, which denote named entities of different types in plain texts. This challenge was addressed in Task 1 of FactRuEval-2016 evaluation.
In the context of this evaluation, our team, acting for the Institute for System Programming of the Russian Academy of Sciences, proposed two approaches to exploiting information, mined from Wikidata and Wikipedia, for improving quality of named entity detection methods. In the first approach word2vec word embeddings, computed on Wikipedia, are used along with basic features in tokens classification. The second approach utilizes both Wikipedia and Wikidata to automatically construct a representative corpus for named entity recognition and classification training. Additionally, Wikidata, treated as a property graph, is used to collect named entity specific word dictionaries.
Our approaches (marked with identifier ‘Orange’ in FactRuEval-2016 organizers’ quality evaluation reports) show up promising results, doing especially good for such well-defined class as person, still being appropriate for detecting named entities of other types as well.

Key words: named entity recognition, wiki, machine learning, word embedding, word2vec

Introduction

Named entity recognition and classification (NERC) is an important information extraction task which is aimed at sifting through plain texts for such information units as human names, locations, organizations, facilities, products, dates, geopolitical entities, holidays and so on.

During FactRuEval-2016 evaluation, Task 1 was fully devoted to classical NERC. Participants' efforts were to be concentrated on detecting entities of three most popular types—person, location, organization. Additionally, some sophistication of the problem was introduced: it was required to distinguish between locations, meaning some geographical place, and locations, meaning an organization or group of people, as in 'Russia is celebrating Victory Day'. The later entity type is referred to as locorg.

Results of our team, acting for Information Systems Group of the Institute for System Programming of the Russian Academy of Sciences, are marked with identifier 'Orange' in FactRuEval-2016 final evaluation reports.

The rest of the article is structured as follows. Related work is discussed in Section 1. Section 2 outlines some important features of wiki-resources, exploited in our work. Section 3 describes our approach to FactRuEval-2016 named entity recognition task. In Section 4 evaluation results are provided. We wrap up with some concluding thoughts in the final section.

1. Related work

One of the earliest approaches to named entity recognition and classification was based on human-defined rules and heuristics (Rau, 1991). For now methods, based on machine learning, seem to be more promising (Zhang & Johnson, 2003), as they are easier to build and adapt to new domains. But they still face a major natural language processing problem—data sparsity: words are represented with numerical vectors of high dimensionality, thus requiring huge train corpus for building a representative model. Initial attempt to cope with data sparsity was building dictionaries of somehow similar words and adding features, indicating whether classified word belongs to one of them (Zhang & Johnson, 2003). Recent approaches utilize language models (Miller et al., 2004)—Brown clusters (Brown et al., 1992), word2vec (Mikolov et al., 2013)—in attempt to reduce problem dimensionality. Another way is to semi-automatically construct more representative training corpus (Nothman et. al, 2013).

In our work we try to compare both approaches: in the first arrangement we bet on word2vec features, computed on Wikipedia; in the second arrangement we stake on training upon automatically created corpus, made up from Wikipedia and Wikidata.

2. Wiki-resources overview

For better understanding of our approach, some Wikipedia and Wikidata features, vital for our work, are briefly described below.

2.1. Wikipedia

Wikipedia (www.wikipedia.org) is a free online encyclopedia, which can be updated and edited by any user. Its large volume and rich link structure make Wikipedia an invaluable resource for solving many natural language processing problems (Milne & Witten, 2008; Ratinov et al., 2011; Turdakov et al., 2014). Wikipedia's textual articles describe entities of the real world, linked with each other to simplify navigation through different parts of the encyclopedia.

2.2. Wikidata

Wikidata (www.wikidata.org) is a free online machine-readable multilingual knowledge base, interlinked with other Wiki resources. Wikidata can be thought of as a property graph (Jouili & Vansteenberghe, 2013), where ontological classes and instances are represented with vertices and relations—with edges. In Wikidata terminology they are called, respectively, items and properties. Items are assigned special identifiers, started with 'Q' and followed by some number; property identifiers start with 'P', followed by some number as well. Both items and properties may have several textual representations on the specified language; some representation is selected as the main one, called label, while others are called aliases. Items may also have links to corresponding Wikipedia articles on different languages.

For example, Wikidata contains such items as 'geographical object' (Q618123, Wikipedia article: 'Geographical feature'), 'fictional location' (Q3895768, Wikipedia article: 'Fictional location', aliases: 'fictional place', 'mythical location', 'legendary place', ...), 'Narnia' (Q2886622, Wikipedia article: 'Narnia (country)'). Sample properties are: 'subclass of' (P279), 'instance of' (P31), 'given name' (P735). One can infer relations between unconnected items with Wikidata graph traversal: 'Narnia' (Q2886622) is an instance of 'fictional location' (Q3895768), as there is an 'instance of'-'subclass of' path through 'fictional country or state' (Q1145276).

3. Method description

Our method is based on sequential traversal and classification of text tokens. Computed token labels are then used to determine named entity type and boundaries in text.

Valid labels are constructed from supported named entity types using BILOU encoding scheme (Uchimoto et al., 2000). Four labels are generated for each type: B-TYPE (for example, B-PERSON, B-ORGANIZATION) is used to indicate entity beginning, I-TYPE indicates token in the middle of the entity, L-TYPE specifies ending of the entity, U-TYPE marks single token entities. Additionally, there is O label, which is assigned to non-entity tokens. In FactRuEval-2016 evaluation the following named entities are supported: person, location, organization and locorg. Sequence of assigned labels can be naturally decoded (Ratinov & Roth, 2009) to restore named entity mentions in the plain text.

To compute correct label, each token is converted into feature vector which is fed into one-vs-rest multiclass linear SVM classifier (Fan et al., 2008). The following groups of basic extractors are used to fill up feature vector of the token: word-level (Zhang & Johnson, 2003), local context and global context extractors (Ratinov & Roth, 2009).

Word-level feature extractors are used to collect pieces of information, sealed in a word itself. We use the following extractors of this group: token affixes of lengths from one to four; token text, part-of-speech tag, lemma and digit normalized token form, where all digits are replaced with a special character; predicates, indicating token properties: starting from capital letter, containing characters of the same class (digits, quotation marks), being constructed only from characters of the same class (digits, digits or letters, non-letters, uppercase letters).

Local context features encode information from nearby area of a certain token. We selected the following features of this group: token position in sentence (being first token; not being first token; not being last token); BILOU labels assigned to up to three previous tokens of the analyzed text.

The final group of features—global context features—tries to exploit information from the whole document or from some pretty large area around the token under consideration. We utilize feature values and labels distributions of tokens, sharing the same (case ignored) form among 200 previous tokens for labels and 2,000-token window around the target token for feature values.

We should also mention, that when building feature vector for a certain token, not only its features are collected—features of up to three surrounding tokens from the same sentence are appended to the vector as well.

In addition to basic, we also use two groups of more sophisticated features, which are described below.

3.1. Dictionary features

Building dictionaries is the first place, where we employed wiki-resources power for NERC task. In our approach Wikidata is used to construct a number of named entity specific dictionaries. For each of them we generate a separate feature, indicating whether it contains token under consideration or not.

To build a set of dictionaries we treat Wikidata as a property graph and select a number of seed items, characterizing current named entity type. For example, for location we select ‘geographical object’ (Q618123), ‘historic site’ (Q1081138), ‘fictional location’ (Q3895768); for organization—‘organization’ (Q43229); for person—‘human’ (Q5), ‘fictional character’ (Q95074). Then we traverse Wikidata graph along ‘subclass of’ edges in reverse direction (from more general to more specific class). The final step is performed along ‘instance of’ edge (in the reverse direction, as well), delivering a number of items, representing named entity instances of the required type (Fig. 1).

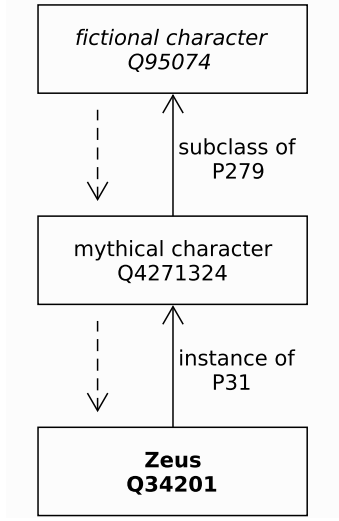


Fig. 1. Direction of Wikidata graph traversal

Labels and aliases of these instances are collected separately. The final step is to encode each entry with BILOU (actually, O is not used) and group similarly marked tokens together. Thus for each named entity type we build eight dictionaries: B, I, L and U for aliases and labels.

Additionally, we collect dictionaries of human names, surnames and nicknames. Names are gathered from vertices, linked with 'female given name' (Q11879590), 'male given name' (Q12308941), 'unisex name' (Q3409032) via 'instance of' edge; surnames—from vertices, linked with 'family name' (Q101352), 'surname' (Q4116295), 'cognomen' (Q777342), 'nomen' (Q15238609). Nicknames are collected as 'nickname' (P1449) value of vertices, denoting humans. Further steps for generating dictionaries are similar to what has already been outlined above.

3.2. Word2vec features

Another facility to utilize wiki-resources power is to build a language model. We extract plain texts from Wikipedia articles (excluding the shortest ones as they tend to be noisy), lemmatize them and compute word embeddings with word2vec implementation (skip-gram model; hierarchical softmax enabled; window size of 10; vector dimensionality of 100). Computed vector elements are added as separate features.

3.3. Training data

When building named entity recognition model we used two separate training sets. For the first arrangement we utilized data, provided by FactRuEval organizers. As the training set was rather small (122 documents), we decided to inject more «knowledge» to the model via adding word2vec features, computed over Wikipedia.

For the second arrangement we chose to exploit only the power of Wikipedia and Wikidata mixture—FactRuEval data was abandoned. Our approach is significantly navigated by (Nothman et. al, 2013), but in contrast it requires less manual markup. Below we provide some details.

The first step here is to derive Wikipedia articles classification into named entity types. This is performed in the same way, as has been described for generating dictionaries: Wikidata graph is traversed from a number of seed vertices, denoting a single named entity type; this type is then assigned to Wikipedia articles, corresponding to each of the collected target vertices.

Further steps repeat the approach, described in (Nothman et. al, 2013): extracting text with links, mapping links to named entities, enriching text with additional mentions, selecting sentences for the train corpus.

We should point out, that our approach to constructing Wikipedia articles classification is less time-consuming, comparing to (Nothman et. al, 2013): we derive required information directly from Wikidata graph traversal instead of manually collecting initial seed set of marked up Wikipedia articles, further used to train a classifier. However, there is some disadvantage: our named entity recognizer fails to tell location from locorg, which is expected in FactRuEval-2016, due to lack of such distinction in the training data.

4. Evaluation results

In this section we present evaluation results for two NERC arrangements: FactRuEval- and Wiki-based.

4.1. FactRuEval-based arrangement

We exploit training corpus, provided by FactRuEval-2016 organizers, to build named entity recognizer, capable of distinguishing entities of the four required types: person, organization, location, locorg. To evaluate aptitude of different feature groups we examine several combinations (Table 1). Contribution of word2vec and dictionary features, used separately, looks significant; still their mixture delivers some final bits to the result.

Table 1. Feature group combinations in FactRuEval-based arrangement

feature set	precision	recall	f1
basic features	0.7357	0.6186	0.6720
basic + dictionary features	0.8098	0.6988	0.7502
basic + word2vec features	0.8093	0.7241	0.7643
basic + dictionary + word2vec features	**0.8257**	**0.7408**	**0.7810**

Table 2 contains detailed evaluation results for all-features arrangement. Our method is good for such well-defined named entity type as person, while it also provides reasonable results for other types.

Table 2. Per-type evaluation results for full feature set FactRuEval-based arrangement

type	precision	recall	f1
person	0.9340	0.8675	0.8995
location	0.7259	0.6944	0.7098
organization	0.7844	0.6548	0.7137
locorg	0.7858	0.7251	0.7542
overall	0.8257	0.7408	0.7810

4.2. Wiki-based arrangement

For training we use texts, generated from Wikidata and Wikipedia, as has been described earlier. Word2vec features, computed on Wikipedia, are not used in this arrangement. On presenting results we ignore difference between location and locorg types, as they are not distinguished in the training documents. Table 3 and Fig. 2 depict FactRuEval evaluation results for different volumes of the training corpus.

Table 3. Influence of training corpus volume on results quality

number of training documents	FactRuEval testset			FactRuEval devset		
	precision	recall	f1	precision	recall	f1
500	0.8430	0.6342	0.7238	0.8190	0.6893	0.7486
1,000	0.8476	0.6274	0.7210	0.8197	0.6826	0.7449
2,000	0.8628	0.6360	0.7323	0.8278	0.6936	0.7548
3,000	0.8655	0.6332	0.7313	0.8267	0.6880	0.7510
4,000	0.8687	0.6353	0.7339	0.8413	0.6918	0.7593
5,000	0.8763	0.6388	0.7389	0.8427	0.6898	0.7586
6,000	0.8734	0.6396	0.7384	0.8486	0.6931	0.7630
8,000	0.8817	0.6470	0.7463	0.8508	0.6927	0.7637
10,000	**0.8819**	**0.6475**	**0.7467**	**0.8525**	**0.6942**	**0.7652**

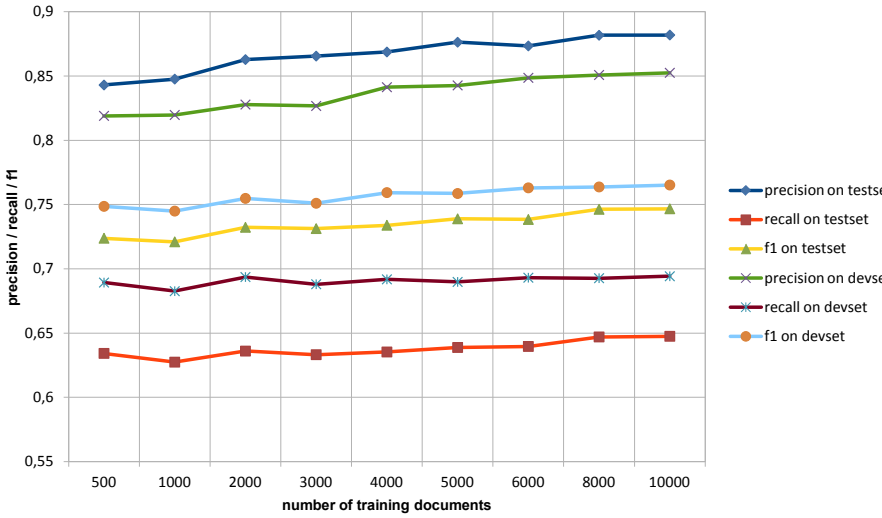


Fig. 2. Influence of training corpus volume on results quality

As one can notice, increasing volume of training corpus promotes evaluation results quality improvement, which starts to slow down at around 6,000–8,000 documents.

Table 4 reports detailed evaluation results for Wiki-based arrangement. Our method works definitely good for person named entity type; it shows promising results for other types, with the only confusing recall for organizations. We explain it with discrepancy of FactRuEval markup rules and Wikipedia linking policy: there seem to be some amount of named entities, which are not typically used as anchor texts in Wikipedia, thus making our method short on detecting them. Analysis shows that our method typically misses informally named entities ("Минфин", "Минюст", "российский газовый монополист", "парижская Третьяковка") and entities in controversial cases ("интернет", "Рунет" are considered organizations; meaning of "австралийское правительство" is not clear: governmental institution or synonym for "the authorities"; "музей Орсе" is considered organization instead of location or facility).

Table 4. Per-type evaluation results for Wiki-based arrangement (10,000 training documents) on FactRuEval-2016 testset

type	precision	recall	f1
person	0.9624	0.7976	0.8723
location	0.7993	0.7470	0.7723
organization	0.8939	0.4327	0.5831
overall	0.8819	0.6475	0.7467

Conclusion

In this paper we presented two approaches for named entity recognition and classification, both showing promising results in FactRuEval-2016 evaluation.

The first approach utilizes rich feature set: word-level, local and global context features, word2vec word embeddings and dictionary features. We showed that dictionary and word2vec features, even used separately, manage to significantly amend results quality. Combining all features together facilitated further improvement.

The second approach illustrates, how to utilize the power of Wikipedia and Wikidata mixture for automatically building a large training corpus for NERC task. Our research has shown that even a few thousand labeled documents can be used to achieve comparable results for detecting named entities of several most popular types: person, location, organization.

References

1. *Brown P. F., deSouza P. V., Mercer R. L., Pietra V. J. D., Lai J. C.* (1992), Class-based n-gram models of natural language, Computational Linguistics. Vol. 18, Issue 4, pp. 467–479.
2. *Fan R.-E., Chang K.-W., Hsieh C.-J., Wang X.-R., Lin C.-J.* (2008), LIBLINEAR: A library for large linear classification, Journal of Machine Learning Research, Vol. 9, pp. 1871–1874.
3. *Jouili S., Vansteenberghe V.* (2013), An Empirical Comparison of Graph Databases, Proceedings of the 2013 International Conference on Social Computing, Alexandria, pp. 708–715.
4. *Mikolov T., Chen K., Corrado G., Dean J.* (2013), Efficient Estimation of Word Representations in Vector Space, available at: http://arxiv.org/pdf/1301.3781.pdf
5. *Miller S., Guinness J., Zamanian A.* (2004), Name Tagging with Word Clusters and Discriminative Training, Proceedings of HLT, Boston, pp. 337–342.
6. *Milne D., Witten I. H.* (2008), Learning to link with wikipedia, Proceedings of the 17th ACM conference on Information and knowledge management (CIKM '08), Napa Valley, pp. 509–518.
7. *Nothman J., Ringland N., Radford W., Murphy T., Curran J. R.* (2013), Learning multilingual named entity recognition from Wikipedia, Artificial Intelligence, Vol. 194, pp. 151–175.
8. *Ratinov L., Roth D.* (2009), Design Challenges and Misconceptions in Named Entity Recognition, Proceedings of the Thirteenth Conference on Computational Natural Language Learning, Boulder, pp. 147–155.
9. *Ratinov L., Roth D., Downey D., Anderson M.* (2011), Local and global algorithms for disambiguation to Wikipedia, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies—Volume 1 (HLT '11), Portland, pp. 1375–1384.
10. *Rau L. F.* (1991), Extracting company names from text, Proceedings. Seventh IEEE Conference on Artificial Intelligence Applications, Miami Beach, pp. 29–32.

11. *Turdakov D., Astrakhantsev N., Nedumov Y., Sysoev A., Andrianov I., Mayorov V., Fedorenko D., Korshunov A., Kuznetsov S.* (2014), Texterra: A Framework for Text Analysis [Texterra: infrastruktura dlya analiza tekstov], Proceedings of the Institute for System Programming of RAS [Trudy ISP RAN], volume 26, Issue 1, pp. 421–438.
12. *Uchimoto K., Ma Q., Murata M. Ozaku H., Isahara H.* (2000), Named entity extraction based on a maximum entropy model and transformation rules, Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Hong Kong, pp. 326–335.
13. *Zhang T., Johnson D.* (2003), A robust risk minimization based named entity recognition system, Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, Edmonton, pp. 204–207.

«АППОЗИЦИОНАЛЬНЫЕ» И «СООПРЕДЕЛЯЮЩИЕ» УСЛОВНЫЕ КЛАУЗЫ: К ВОПРОСУ О ЛОКАЛИЗАЦИИ УСЛОВНОЙ СЕМАНТИКИ[1]

Тискин Д. Б. (daniel.tiskin@gmail.com)

СПбГУ, Санкт-Петербург, Россия

В работе демонстрируется семантическая неоднозначность сочинения условных клауз в русском языке, которые могут получать не только «параллельное» прочтение (‘при *A* и при *B*’), но и «соопределяющее» (‘в тех случаях *A*, которые являются и случаями *B*’). Предлагается формальный анализ.

Для аппозициональных чтений известен анализ, согласно которому сама по себе ИГ обозначает свойство, а не индивид или обобщённый квантор. Значение индивида или квантора она приобретает благодаря правилам сдвига типа. Применяя данное рассуждение к условным клаузам, мы заключаем, что *если* не может быть ответствен за референциальный компонент семантики условных клауз — их способность обозначать некоторое множество миров. Учитывая аргументы Д. Льюиса и др. против трактовки условного союза как квантора по мирам, мы заключаем, что *если* практически лишён собственной семантики и лишь определяет, что квантификация осуществляется именно по возможным мирам, а не по моментам времени (как в случае *когда*). Предложенный анализ может быть использован при создании правил автоматического выделения и анализа подобных конструкций.

Ключевые слова: сложное предложение, именная группа, условные клаузы, безвершинные относительные клаузы, конъюнкция, сдвиг типа

“APPOSITIONAL” AND “CO-DETERMINATIVE” CONDITIONAL CLAUSES: IN SEARCH OF THE LOCUS OF CONDITIONALITY

Tiskin D. B. (daniel.tiskin@gmail.com)

St. Petersburg State University, St. Petersburg, Russia

[1] Автор выражает благодарность М. Крижу, О. В. Митрениной, О. А. Митрофановой и трём анонимным рецензентам. Оставшиеся недостатки работы лежат целиком на совести автора.

The paper provides evidence for the claim that the Russian conditional conjunction *esli* 'if' is itself devoid of either conditional or determiner semantics. The argument proceeds as follows. I demonstrate that with conjoined conditionals, just like with some NPs/DPs and with free relatives, one gets not only the immediately obvious "parallel" reading ('for *A* and for *B*') but also the "co-determinative" reading ('for those *A* which are also *B*'). The sort of reading identified in the literature as "appositional" turns out to be a subclass of co-determinative readings.

It has been proposed that appositional readings for NPs/DPs result from the fact that the pertinent DPs denote properties, whereas their conversion into referring or quantificational expressions is performed by type-shifting rules. Applying the same technique to conditionals, I conclude that the conditional *esli* cannot have the semantics of the definite determiner in the domain of possible worlds. Given the influential view (Lewis etc.) that *if* does not quantify over worlds either (that work done rather by "adverbs of quantification", which may be overt, e. g. *always* or *usually*, or covert), *esli* ends up free from any semantic duty, except that—as I argue—it determines whether the quantification over worlds or over time instances takes place (cf. *esli* vs. *kogda* 'when').

The proposed analysis may be used as guidance for the development of automatic recognition and analysis rules for such constructions.

Keywords: compound sentences, NP, DP, conditional clauses, free relatives, conjunction, type shift

> A computer scientist earns brownie points for showing that two things that seem different are actually the same, whereas a linguist earns brownie points for showing that two things that seem the same are actually different.
>
> Chung-chieh Shan

1. Аппозициональные и соопределяющие прочтения

Сочинённые именные группы[2] могут, наряду с обычным прочтением ('верно для *A* и верно для *B*'), иметь прочтение, при котором обе ИГ характеризуют один и тот же объект (индивид или множество). Поэтому они не сочетаются с *оба* и не вызывают согласования глагола по множественному числу, если только каждая из них не является множественной, как в (1).

---

[2] Мы используем термин *именная группа* (ИГ) как родовой, а термины *NP* и *DP* как видовые и относящиеся к случаям, когда вопрос о наличии в ИГ детерминатора решается однозначно.

(1) *Если* [*отцы-основатели США*]$_i$ *и* [*авторы американской конституции*]$_i$ *первоначально даже не задумывались о такой проблеме, как возможность выхода военных из-под опеки гражданского общества…*[3]

Такие конъюнкции называют *аппозициональными* (appositional). Чтобы охарактеризовать привычные прочтения для конъюнкций, мы будем пользоваться термином *параллельные*.

На английском материале установлено, что аппозициональные прочтения существуют у определённых и неопределённых ИГ, у ИГ с посессором (Hoeksema 1988: 36, Coppock & Beaver 2015: 382); Chaves (2007: 80) обнаруживает у сочинённых DP с универсальным детерминатором *every*, кроме параллельного, ещё и прочтение, при котором значением конъюнкции является **пересечение** множеств, к которым отсылают предикаты-рестрикторы (*husband* и *father* в (2)):

(2) *Every husband and every father is required to do this test.*
'Всякий муж и (всякий) отец должен пройти этот тест.'

Такие прочтения мы будем называть *соопределяющими*, поскольку множество объектов, характеризуемых конъюнцией, по-своему ограничивается каждым из конъюнктов, так что ни один не может быть опущен без изменения смысла. При множественности каждого из конъюнктов аппозициональная интерпретация в узком смысле (полное совпадение значений конъюнктов) отличается от соопределяющей (значение конъюнкции — пересечение значений конъюнктов).

Так или иначе, русские аналоги (2) не имеют аппозициональных и соопределяющих прочтений. Собственным именам (Hoeksema 1988) и DP, возглавляемым *some* 'какой-то/некоторые' (Coppock & Beaver 2015: 382), в таких прочтениях также отказано.

На английском материале нам не удалось найти упоминаний в литературе аппозициональных (или соопределяющих) прочтений для конъюнкций ИГ во множественном числе, однако в русском языке, где употребление лексического определённого детерминатора необязательно, такие примеры отыскать можно, хотя они и нечасты (ср. (1)).

1.1. Соопределяющие прочтения для клауз с вершиной

В русском языке соопределяющие прочтения возможны для сочинённых клауз. Так, для обычных рестриктивных клауз с ненулевой вершиной Лютикова (2008) отмечает, что при конъюнкции значение пересечения множеств (в нашей терминологии — соопределяющее) доступно всегда, а значение

[3] Если не указано иное, примеры найдены в сети Интернет с помощью поисковой системы Google; примеры, помеченные «НКРЯ», взяты из Национального корпуса русского языка (http://ruscorpora.ru).

объединения множеств (параллельное) затруднено или невозможно (примеры самой Е. А. Лютиковой):

(3) *Давайте встретимся с избирателями, которые живут в этом районе и которые поддерживают «Яблоко».* [ОК‘и живёт, и поддерживает’; ? ‘живёт или поддерживает’]

Если значение пересечения оказывается невозможно по сторонним причинам (получалось бы противоречивое свойство), предложение не имеет удовлетворительной интерпретации (также пример Е. А. Лютиковой):

(4) *# Давайте встретимся с избирателями, которые проголосовали за Зюганова и которые не пришли на выборы.*

Если же вершина у клауз формальная (*то*), соопределяющая интерпретация практически необходима[4]; соответственно, она практически невозможна при наличии **повтора** *то*.

При допущении, что относительные придаточные с вершиной обозначают свойства (семантический тип ⟨*e*,*t*⟩) и семантически сочетаются со значением вершины посредством модификации предиката (*Predicate Modification*, см. Heim & Kratzer 1998: 65), эти эффекты легко объяснимы: конъюнкция свойств и должна давать свойство, влекущее оба исходных, а значение связанного с вершиной детерминатора (если он есть) сочетается уже с этим составным свойством.

Впрочем, наряду с соопределяющими для клауз с вершиной возможны и прочтения, аналогичные аппозициональным[5], — если обе относительные клаузы очерчивают одно и то же множество объектов (либо же первая очерчивает собственное подмножество множества, очерчиваемого второй). Тогда бывает возможно добавить во вторую клаузу *поэтому*[6]:

(5) *Приведем некоторые процессы распада, при которых сохраняется барионное число и которые поэтому разрешены (и наблюдались)…*

1.2. Сочинение безвершинных относительных клауз

Несколько иначе ведут себя *безвершинные* относительные клаузы (free relatives, БОК). Если отбросить примеры типа (6), где значением вложенной клаузы

---

[4] Ср., однако, в НКРЯ: *Команда исходно была разделена на две части: те, кто играет и кто присутствует*. О том, следует ли здесь видеть эллипсис второго *те*, см. Kaplan (2008).

[5] Для предикатов, каковыми семантически являются и клаузы с вершиной, в значении ‘параллельное’ используется термин *split*, а в значении ‘аппозициональное/соопределяющее’ — термин *joint* (Heycock & Zamparelli 1999).

[6] Ср. анализ *therefore* ‘поэтому’ как анафорического элемента у Brasoveanu (2010). Мы благодарны М. Крижу за указание на статью А. Брасовяну.

является вопрос (Groenendijk & Stokhof 1984; см. также «дизъюнктивные» БОК в Ross 2002), они достаточно легко допускают и соопределяющие прочтения (7), и параллельные (8–9), хотя последние возможны, может быть, почти исключительно в идиоматизированных случаях или когда соопределяющее прочтение привело бы к появлению противоречивого свойства (ср. (4)). Случаи, когда возможны (или явно предпочтительны) только аппозициональные прочтения, тоже существуют (10).

(6) *Никто не знает, что было и что будет…*

(7) *От этой еды у меня случалась изжога, возникала отрыжка, но я ни с чем не спорил, я ел, что мне дают и что вызывало одобрение Саломеи.* [НКРЯ]

(8) *Высоцкий гражданином был… И говорил, что надо и не надо.*

(9) *Я видел еще массу всего в промежутках, что я мог понимать и что не мог.*

(10) *Каждый подарок походил на мастер-класс, когда всяк потчевал всякого, чем рад и чем богат!* [НКРЯ]

2. Семантика сочинённых условных клауз

Как и БОК, условные клаузы свободно допускают соопределяющие прочтения, т. е. прочтения, при которых условием выполнения консеквента является **совместное** выполнение обоих условий, названных антецедентами:

(11) *Если же это правда и если это долго не изменится, — я не могу себе представить выше счастья.* [НКРЯ]

(12) *Воздействие одного человека — если он соблюдает правила и если они адекватны для данной экосистемы — практически неощутимо.* [НКРЯ]

Параллельные прочтения также возможны[7] (как в случаях, когда иначе получилось бы противоречие (14), так и в иных (13)):

(13) *При описании наблюдений следует избегать категоричности выводов, если число наблюдений недостаточно и если в наблюдениях имеются неясности или сомнения в точности наблюдений…* [НКРЯ]

(14) *Если да и если нет, позвоните нам.* [НКРЯ]

---

[7] Параллельное чтение ещё может быть получено **дизъюнкцией** антецедентов: *Съемное протезирование зубов применяется, если пациенту противопоказана имплантация или если есть другие причины отказа от операции.*

Предпочтение соопределяющих прочтений явственно выступает в присутствии в главной клаузе лексем, квантифицирующих по возможным мирам, выполняющим антецедент условной конструкции (см. ниже об устранителях гомогенности):

(15) *Если получиться* [sic] *и если Вадим не будет против то непременно в следующем году поедим* [sic] *туда же.*

(16) *Вот только минуты скорее всего не хватит, если нет навыка и если все сильно запущено.*

Аппозициональные прочтения для условных клауз возможны, хотя и достаточно редки. Они относительно продуктивны в устойчивом выражении *если я ошибаюсь и* (*если*)… (17), но встречаются и в других случаях (ср. одно из прочтений (18)). Как и в случае аппозициональных прочтений для определённых и неопределённых DP, обозначающих один и тот же индивид, тут обе клаузы описывают одно и то же множество миров (ср. von Fintel 2011: 1520 ff.), тогда как в случае соопределяющих прочтений они описывают множества пересекающиеся и вместе обозначают пересечение.

(17) *Но если я ошибаюсь и если такая встреча произойдет, мы обсудим ее предварительно на Лиге избирателей…*

(18) *А что если самореализация не состоялась и если часть задач осталась нерешенной?*

3. Анализ поведения условных клауз

3.1. Проблема и путь решения

В ряде работ (см. von Fintel 2011: 1526) были отмечены сходства между условными клаузами и БОК. В теориях, где такое сходство объявляется неслучайным, *если* может отводиться (Schlenker, 2004) роль оператора определённой дескрипции. БОК тоже анализируются как определённые дескрипции, в т.ч. потому, что проявляют свойство *максимальности*[8]: так, в (10) каждая из БОК обозначает **максимальный** множественный индивид (Link 1983),

[8] Križ (2014, 2015), соответственно, указывает на сходство в **немаксимальных** интерпретациях между условными клаузами и определёнными дескрипциями.

удовлетворяющий называемому в клаузе свойству[9], т. е. сумму **всего**, чем богат или чем рад угостить данный гостеприимец.

Križ (2015) утверждает, что сходство между условными клаузами и определёнными дескрипциями проявляется в том, что как те, так и другие обладают свойством *гомогенности*. Это означает, что наличие исключений не делает всё предложение ложным (но делает «не вполне истинным»). Для определённых множественных DP исключение — это индивид, являющийся частью значения дескрипции, но не выполняющий предицируемого свойства (Смит в (19)); для условных клауз исключение — возможный мир, выполняющий антецедент, но не выполняющий консеквента (любой из тех «маловероятных» миров, где Мэри приходит вместе с Джоном, в (20); оба примера принадлежат М. Крижу).

(19) *The professors smiled.*
'Профессора улыбнулись.'
Смит, один из профессоров, никогда не улыбается. Он не улыбнулся и в этот раз, но другие улыбнулись. Предложение не рассматривается как ложное.

(20) *If Mary comes, Sue will be pleased.*
'Если Мэри придёт, то Сью будет обрадована.'
Сью устраивает вечеринку, на которой будет рада видеть Мэри. Джон вряд ли придёт, но если придёт, то испортит вечеринку. Предложение не рассматривается как ложное, хотя вероятность того, что Джон тоже придёт, не равна нулю (пусть и мала).

Гомогенность может быть устранена. В случае DP её устранителем является *all* 'все'; при его добавлении в (19) предложение становится ложным в описанной ситуации. Для условных клауз гомогенность может быть устранена словами типа (*not*) *necessarily* '(не) необходимо, (не)обязательно'. Действительно, насчёт (20) нельзя сказать, что Сью **обязательно** порадует появление Мэри, потому что в случае появления Мэри и Джона вместе радости не будет. Налицо сходство между определённым детерминатором *the* и условным союзом *if*.

Наши данные ставят тезис «*если* = 'the'» под сомнение, по крайней мере для русского языка: если конъюнкция оперирует уже двумя готовыми дескрипциями, результатом её применения должно быть параллельное прочтение: 'если *p*, то *q*, да и если *r*, тоже *q*'; то же наблюдалось бы и для определённых и неопределённых ИГ в английском языке, если бы артикль всегда нёс семантику детерминатора.

---

[9] Впрочем, как отмечает Hinterwimmer (2008), БОК в германских языках регулярно полисемичны: кроме максимальной интерпретации, доступна бывает ещё экзистенциальную интерпретацию, как в немецком примере В. Штернефельда: *Wer nimmt, was ihm nicht gehört, ist ein Dieb* 'Кто берёт то [= что-то из того], что ему не принадлежит, — вор' (Sternefeld 2005). Примером немаксимального чтения для конъюнкции может, по-видимому, служить (7).

Получить соопределяющие прочтения ('если разом *p* и *r*, то *q*') можно, если считать, что *если* ничего в семантику не вносит, но за максимальность отвечает либо некоторый нулевой элемент выше в структуре, чем *если*, либо лексически не выраженная семантическая операция (сдвиг типа). Если это нулевой элемент, то соображения экономии нулей (нулевых единиц не следует порождать больше, нежели необходимо, чтобы предложение было грамматичным и имело интерпретацию) могут помочь объяснить редкость параллельных прочтений для условных клауз. Аналогично можно поступить и с БОК, для которых иногда постулируется нулевая вершина (Никунласси 2008; аналогично, Зализняк и Падучева (1975) говорят об эллипсисе местоимения в главной клаузе; см. критическое обсуждение этой точки зрения в статье Лютикова (2015: 78 и сл.)).

Для условного союза идея несемантичности — в том смысле, что *если* не соединяет двух клауз, а присоединяет аргумент к (иногда нулевому) наречию-квантору вроде *всегда* или *как правило*, — возводится к Lewis (1975) и Kratzer (1991) (но см. критику такого трёхкомпонентного подхода в Huitink (2007)). Как сохранить аналогию между условными клаузами и определёнными дескрипциями при допущении десемантизированности *если*?

3.2. Уточнение решения: сдвиг типа и значение если

Coppock & Beaver (2015: 382) используют аппозициональные прочтения как аргумент в пользу того, что сама по себе (определённая или неопределённая) DP, даже если имеет лексически выраженный детерминатор, может обозначать свойство (а не индивид или обобщённый квантор)[10]. Действительно, в таком случае аппозициональная семантика следует из возможности применить правило *сдвига типа* (type shift; Partee 1987) iota как к каждому из конъюнктов в отдельности, так и к конъюнкции в целом. Эта операция сопоставляет свойству (значению предиката) единственный индивид, обладающий данным свойством (если таковой имеется).

$$\text{IOTA} : \lambda x.P(x) \rightsquigarrow \iota X.P(X)$$

Для определённых дескрипций и БОК результатом применения IOTA будет единственный (максимальный множественный) индивид *X*, удовлетворяющий дескрипции или, соответственно, обозначаемому клаузой свойству; для условных клауз результатом применения IOTA будет максимальный индивид, составленный из миров, выполняющих обозначаемое клаузой условие. (Для (16) в его соопределяющем прочтении, к примеру, это миры, где нет навыка

[10] С синтаксической стороны аналогичный тезис выдвигает Winter (2001: 178), утверждающий, что аппозициональная конъюнкция происходит на уровне NP, но не может происходить на уровне D′ или DP (откуда запрет на $*some_i\ P\ and\ some_i\ Q$), поскольку только на этом уровне добавляется признак числа, так что конъюнкция двух NP может получить признак [+ *sg*] (~ обозначать один объект, а не два).

и где всё сильно запущено; таким образом, ιοτα применяется уже к конъюнкции, а не к каждому из конъюнктов в отдельности, как в (14).)

Хотя решение, при котором *если* не несёт ни условной семантики, ни семантики определённости, тогда как эти функции выполняются фонологически нулевым квантором и операцией сдвига типа, трудно назвать экономным, лучшее решение предложить непросто. Помимо этого, мы предполагаем, что *если* всё же не совсем пуст. Чтобы это продемонстрировать, укажем на проблематичность (21), хотя ср. приемлемый пример (22), принадлежащий анонимному рецензенту:

(21) *Мы напишем вам, если освободится место* {?*или* / ??*и*} *когда будет открыт набор.*

(22) *Когда место освободится и если квалифицированных претендентов на него будет немного, мы вам позвоним.*

Так же как и *если*, *когда* позволяет гомогенность и её устранение (хотя, по-видимому, отличается от *если* тем, что и в присутствии устранителя, и без него допускает как параллельные, так и соопределяющие прочтения, причём параллельные возможны не только в тех случаях, когда соопределяющие давали бы противоречие — пустое множество миров)[11]:

(23) *Разговаривайте с малышом* (*всегда*)*, когда он видит вас и когда вы подходите к нему* [соопределяющее].

(24) *Я была виновата всегда, когда он замахивался и когда материл и обзывал дурой* [оба прочтения].

(25) *Никакие серьезные разговоры с мужчиной нельзя вести, когда он устал и когда он голоден* [оба прочтения].

(26) *Дождь был сочувствующий — он почти всегда прекращался, когда мы ставили лагерь и когда снимали его* [параллельное].

(27) *Это связано с большим количеством ошибок врачей, которые назначают антибиотики, когда нужно и когда не нужно* [параллельное].

Более того, устранители гомогенности наподобие *обычно* или *всегда* годятся и для *если*, и для *когда*. Тем не менее, как мы видим в (21), конъюнкция и дизъюнкция *если… и когда…* менее приемлема, чем в случае *если… если* или *когда… когда*. Причину этого мы предлагаем видеть в семантическом различии между *если* и *когда*: клауза с *если* обозначает множество возможных миров, а клауза *когда* — множество временных точек. Фонологически нулевой

[11] Здесь есть, впрочем, трудность, связанная с тем, что устранитель гомогенности является одновременно и вершиной для клауз с *когда*.

или выраженный лексически квантор (*всегда*, *часто*, ср. специфический для *когда* квантор *всякий раз*) и сдвиг типа, преобразующий свойство в максимальный индивид, одни и те же для клауз с *если* и клауз с *когда*, т. е. они в известных пределах безразличны к семантическому типу аргумента.

4. Выводы

Выше мы продемонстрировали, что в русском языке конъюнкция условных клауз может интерпретироваться несколькими способами: как объединение множеств возможных миров, обозначаемых конъюнктами (параллельные прочтения); как пересечение этих множеств (соопределяющие прочтения); наконец, как одно и то же множество, названное дважды, или как подмножество и объемлющее его множество (аппозициональные прочтения — подкласс соопределяющих, допустимый, когда сочинённые элементы сообщают дополнительную информацию о референте, которую слушающий не использует для его идентификации; ср. термин *apposition* ‘приложение’). Возможно, причина допустимости в русском языке соопределяющих (и аппозициональных) прочтений для условных клауз та же, что и для ИГ во множественном числе (как в (1)), и связана с большей по сравнению с «артиклевым» английским языком активностью сдвигов типа. В то же время, для ИГ даже в русском языке такие прочтения редки (и, видимо, лексически ограниченны, ср. Chaves 2007: 81), тогда как для условных клауз они не менее частотны, чем параллельные. Многие вопросы о (степени) приемлемости тех или иных прочтений, затронутые в настоящей работе, требуют количественных исследований.

Сделанные в настоящей статье наблюдения могут найти применение при разработке систем автоматического анализа языка на основе правил (rule-based methods). В синтаксическом отношении мы не предлагаем никакого пересмотра структуры условных клауз, за исключением известной со времён Lewis (1975) идеи о наличии в структуре сложного предложения с условным придаточным невыраженного квантора: [[Q [*если* S_1]] [(*то*) S_3]]. Семантическое различие, с нашей точки зрения, обусловливается тем, на каком этапе применяется правило сдвига типа. Соответственно, анализатор не может определить, имеет ли он дело с параллельным чтением или же с соопределяющим, исходя только из лексического наполнения предложения. Можно, однако, сформулировать ряд правил, которыми следует руководствоваться при автоматической семантической интерпретации.

1. Для сочинительных союзов (на примере *и*) следует ввести по две записи в лексиконе, соответствующих обычной булевой конъюнкции (u_1) и образованию мереологической суммы (u_2).
2. Если сочинённые условные клаузы обозначают несовместимые положения дел (ср. (14) для *если*, а также (26–27) для *когда*), из двух омонимов для конъюнкции следует выбирать суммирующую конъюнкцию $[\![u_2]\!] = \lambda X\, \lambda Y\, .X \oplus Y$, моделирующую параллельное прочтение (ср.

Link 1983); *X* и *Y* обозначают множественные индивиды, составленные из миров.

3. Если одна из клауз обозначает подмножество миров, обозначаемых другой, следует выбирать булеву конъюнкцию $[\![u_1]\!] = \lambda p\ \lambda q\ \lambda w.\ p(w) \wedge q(w)$, моделирующую соопределяющее прочтение (ср. (19–20), демонстрирующие этот эффект, при очевидной логической возможности суммирующей интерпретации, которая, однако, не наблюдается в русском языке).
4. Если не выполнено ни одно из условий 2–3, следует констатировать неоднозначность, моделируемую как невозможность разрешить омонимию конъюнкции.
5. Поскольку сама по себе клауза всегда обозначает пропозицию, а *Q* всегда требует аргумента — мереологической суммы, в случае 2 сдвиг типа iota применяется до конъюнкции, а в случае 3 — после конъюнкции.

Литература

1. *Зализняк А. А., Падучева Е. В.* (1975). К типологии относительного предложения. Семиотика и информатика, вып. 6, с. 51–101.
2. *Лютикова Е. А.* (2008). Загадки русских относительных предложений. Материалы к докладу на конференции «Синтаксические структуры — 2».
3. *Лютикова Е. А.* (2015). Безвершинные относительные предложения в русском языке: эмпирические данные и теоретические проблемы. Вестник МГГУ им. Шолохова. Сер. Филологические науки, № 3, с. 74–85.
4. *Никунласси А.* (2008). Приместоименно-относительные конструкции в современном русском языке. PhD thesis, University of Helsinki.
5. *Brasoveanu A.* (2010). Decomposing modal quantification. Journal of Semantics, vol. 27, no. 4, pp. 437–527.
6. *Chaves R. P.* (2007). Coordinate structures: Constraint-based syntax-semantics processing. PhD thesis, University of Lisbon.
7. *Coppock E., Beaver D.* (2015). Definiteness and determinacy. Linguistics and Philosophy, vol. 38, pp. 377–435.
8. *von Fintel K.* (2011). Conditionals. In K. von Heusinger, C. Maienborn and P. Portner (eds.), Semantics: An international handbook of meaning, vol. 2, pp. 1515–1538.
9. *Groenendijk J., Stokhof M.* (1984). Studies on the semantics of questions and the pragmatics of answers. PhD thesis, University of Amsterdam.
10. *Heim I., Kratzer A.* (1998). Semantics in generative grammar. Blackwell.
11. *Heycock C., Zamparelli R.* (1999). Friends and colleagues: Plurality and NP-coordination. In Proceedings of NELS 30, pp. 341–352.
12. *Hinterwimmer S.* (2008). Why free relatives sometimes behave as indefinites. In Proceedings of SALT, vol. 18, pp. 411–428.
13. *Hoeksema J.* (1988). The semantics of non-Boolean “and”. Journal of Semantics, vol. 6, pp. 19–40.
14. *Huitink J.* (2007). How to unify restrictive and conditional *if*-clauses. In Proceedings of the 16th Amsterdam Colloquium, pp. 115–120.

15. *Kaplan A.* (2008). The proper role of movement and ellipsis in discontinuous coordination. In Proceedings of WCCFL 26, pp. 297–305.
16. *Kratzer A.* (1991). Conditionals. In A. von Stechow and D. Wunderlich (eds.), Semantik: ein internationales Handbuch der zeitgenössischen Forschung, pp. 651–656.
17. *Križ M.* (2014). Conditionals, monotonicity, and definite descriptions. SPE 7 poster.
18. *Križ M.* (2015). Aspects of homogeneity in the semantics of natural language. PhD thesis, University of Vienna.
19. *Lewis D.* (1975). Adverbs of quantification. In E. L. Keenan (ed.), Formal semantics of natural language. Cambridge University Press, pp. 3–15.
20. *Link G.* (1983). The logical analysis of plurals and mass terms: A lattice-theoretical approach. In R. Bäuerle, C. Schwarze and A. von Stechow (eds.), Meaning, use and interpretation of language. De Gruyter, Berlin.
21. *Partee B.* (1987). Noun phrase interpretation and type-shifting principles. In J. Groenendijk, D. de Jongh and M. Stokhof (eds.), Studies in discourse representation theory and the theory of generalized quantifiers, pp. 115–141.
22. *Ross J. R.* (2002). Conjunctive and disjunctive *wh*-clauses. URL: http://www-personal.umich.edu/~jlawler/CQ-DQ-New.pdf (accessed on 27.01.2016).
23. *Schein B.* (2003). Adverbial, descriptive reciprocals. Philosophical Perspectives, vol. 17, no. 1, pp. 333–367.
24. *Schlenker P.* (2004). Conditionals as definite descriptions. Research on language and computation, vol. 2, no. 3, pp. 417–462.
25. *Sternefeld W.* (2005). Do free relative clauses have quantificational force? In H.-M. Gärtner, S. Beck, R. Eckardt, R. Musan & B. Stiebels (eds.), Between 40 and 60 Puzzles for Krifka. URL: http://www.zas.gwz-berlin.de/fileadmin/material/40-60-puzzles-for-krifka/pdf/sternefeld.pdf (accessed on 27.01.2016).
26. *Winter Y.* (2001). Flexibility principles in Boolean semantics: The interpretation of coordination, plurality, and scope in natural language. MIT Press.

References

1. *Brasoveanu A.* (2010). Decomposing modal quantification. Journal of Semantics, vol. 27, no. 4, pp. 437–527.
2. *Chaves R. P.* (2007). Coordinate structures: Constraint-based syntax-semantics processing. PhD thesis, University of Lisbon.
3. *Coppock E., Beaver D.* (2015). Definiteness and determinacy. Linguistics and Philosophy, vol. 38, pp. 377–435.
4. *von Fintel K.* (2011). Conditionals. In K. von Heusinger, C. Maienborn and P. Portner (eds.), Semantics: An international handbook of meaning, vol. 2, pp. 1515–1538.
5. *Groenendijk J., Stokhof M.* (1984). Studies on the semantics of questions and the pragmatics of answers. PhD thesis, University of Amsterdam.
6. *Heim I., Kratzer A.* (1998). Semantics in generative grammar. Blackwell.
7. *Heycock C., Zamparelli R.* (1999). Friends and colleagues: Plurality and NP-coordination. In Proceedings of NELS 30, pp. 341–352.

8. *Hinterwimmer S.* (2008). Why free relatives sometimes behave as indefinites. In Proceedings of SALT, vol. 18, pp. 411–428.
9. *Hoeksema J.* (1988). The semantics of non-Boolean "and". Journal of Semantics, vol. 6, pp. 19–40.
10. *Huitink J.* (2007). How to unify restrictive and conditional *if*-clauses. In Proceedings of the 16th Amsterdam Colloquium, pp. 115–120.
11. *Kaplan A.* (2008). The proper role of movement and ellipsis in discontinuous coordination. In Proceedings of WCCFL 26, pp. 297–305.
12. *Kratzer A.* (1991). Conditionals. In A. von Stechow and D. Wunderlich (eds.), Semantik: ein internationales Handbuch der zeitgenössischen Forschung, pp. 651–656.
13. *Križ M.* (2014). Conditionals, monotonicity, and definite descriptions. SPE 7 poster.
14. *Križ M.* (2015). Aspects of homogeneity in the semantics of natural language. PhD thesis, University of Vienna.
15. *Lewis D.* (1975). Adverbs of quantification. In E. L. Keenan (ed.), Formal semantics of natural language. Cambridge University Press, pp. 3–15.
16. *Link G.* (1983). The logical analysis of plurals and mass terms: A lattice-theoretical approach. In R. Bäuerle, C. Schwarze and A. von Stechow (eds.), Meaning, use and interpretation of language. De Gruyter, Berlin.
17. *Lyutikova E. A.* (2008). The puzzles of Russian relative clauses [Zagadki russkikh otnositel'nykh predlozhenij]. Talk at the conference "Syntactic Structures 2".
18. *Lyutikova E. A.* (2015). Headless relative clauses in Russian: empirical data and theoretical problems [Bezvershinnye otnositel'nye predlozhenija v russkom jazyke: èmpiricheskie dannye i teoreticheskie problemy]. Vestnik MGGU im. Sholokhova, no. 3, pp. 74–85.
19. *Nikunlassi A.* (2008). Adnominal relative constructions in contemporary Russian [Primestoimenno-otnositel'nye konstrukcii v sovremennom russkom jazyke]. PhD thesis, University of Helsinki.
20. *Partee B.* (1987). Noun phrase interpretation and type-shifting principles. In J. Groenendijk, D. de Jongh and M. Stokhof (eds.), Studies in discourse representation theory and the theory of generalized quantifiers, pp. 115–141.
21. *Ross J. R.* (2002). Conjunctive and disjunctive *wh*-clauses. URL: http://www-personal.umich.edu/~jlawler/CQ-DQ-New.pdf (accessed on 27.01.2016).
22. *Schein B.* (2003). Adverbial, descriptive reciprocals. Philosophical Perspectives, vol. 17, no. 1, pp. 333–367.
23. *Schlenker P.* (2004). Conditionals as definite descriptions. Research on language and computation, vol. 2, no. 3, pp. 417–462.
24. *Sternefeld W.* (2005). Do free relative clauses have quantificational force? In H.-M. Gärtner, S. Beck, R. Eckardt, R. Musan & B. Stiebels (eds.), Between 40 and 60 Puzzles for Krifka. URL: http://www.zas.gwz-berlin.de/fileadmin/material/40-60-puzzles-for-krifka/pdf/sternefeld.pdf (accessed on 27.01.2016).
25. *Winter Y.* (2001). Flexibility principles in Boolean semantics: The interpretation of coordination, plurality, and scope in natural language. MIT Press.
26. *Zaliznyak A. A., Paducheva E. V.* (1975). Towards a typology of relative clauses [K tipologii otnosinel'nogo predlozhenija]. Semiotika i informatika, vol. 6, pp. 51–101.

COREFERENCE IN RUSSIAN ORAL MOVIE RETELLINGS (THE EXPERIENCE OF COREFERENCE RELATIONS ANNOTATION IN "RUSSIAN CLIPS" CORPUS)[1]

Toldova S. Yu. (stoldova@hse.ru),
Bergelson M. B. (mbergelson@hse.ru),
Khudyakova M. V. (mkhudyakova@hse.ru)

National Research University "Higher School of Economics", Moscow, Russia

The work deals with adapting the Russian coreference corpus RuCor annotation system (used for written Russian) to the corpus of Russian oral narratives from the Russian Clinical Pear Stories Corpus (Russian CliPS) (Khudyakova et al., 2016). Russian CLiPS is a corpus of Russian "Pear stories" movie (Chafe, 1980) retellings in clinical populations as compared to neurologically healthy people. The analysis deals with 11 texts by healthy people and 9 texts by people with various types of aphasia. The focus is on the specificity of reference choice in oral retellings and the parameters to be used for the annotation procedure to register deviations in referential choice in spoken discourse as compared to the written one. The specific features for annotation of referential choice in clinical populations are also under discussion. The main claims are as follows. Certain types of speech disfluencies should be integrated into the coreference annotation scheme. These are noun phrases, which are repetitions of a previous referent mention, referent renaming, or name correction. Such occurrences can influence the referent activation; on the other hand, they could shed some light on the process of the referential expression choice. The NP morphosyntactic structure and zero-anaphora should have more granulated set of features for coreference devices, as they are more diverse in spoken discourse. Moreover, certain structures, such as adjectives postposition etc. and some types of zeros are characteristic of referential expressions in spoken discourse.

Keywords: coreference, oral retellings, coreference corpus annotation

[1] The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2016 (grant №16-05-0024) and supported within the framework of a subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the Global Competitiveness Program.

КОРЕФЕРЕНТНЫЕ ОТНОШЕНИЯ В РУССКИХ УСТНЫХ ПЕРЕСКАЗАХ (ИЗ ОПЫТА РАЗМЕТКИ КОРЕФЕРЕНТНЫХ ОТНОШЕНИЙ В КОРПУСЕ «RUSSIAN CLIPS»)

Толдова С. Ю. (stoldova@hse.ru),
Бергельсон М. Б. (mbergelson@hse.ru),
Худякова М. В. (mkhudyakova@hse.ru)

Национальный исследовательский университет
Высшая школа экономики, Москва, Россия

Статья посвящена опыту разметки кореферентных связей в корпусе устных пересказов Russian CliPS (Khudyakova et al., 2016). Корпус представляет собой пересказ фильма о грушах (Chafe, 1980). В статье представлен анализ параметров, которые необходимо учитывать при разметке такого рода текстов. В результате анализа данных, мы предлагаем подходить к разметке кореферентных связей в устных текстах с позиции взаимодействия разных систем: собственно кореферентных цепочек в нарративе, элементов речевых сбоев (например, случаев переименования референта и др.), а также элементов интеракции (например, оценка говорящим степени уверенности в выбранной номинации).

Ключевые слова: кореферентные отношения, устные пересказы, кореферентная аннотация корпуса

1. Introduction

When producing a coherent text, a speaker can use different linguistic devices (NPs) to name an entity (referent): full NPs (*a man, the man with the goat, that man*), anaphoric pronouns (*he, his*), and zero pronouns. When comprehending the text, the listener must make a decision regarding whether a certain NP introduces a new referent in the discourse or relates to a previously mentioned referent. Establishing coreference relations in discourse is a complex process, which depends on various cognitive, discourse and grammatical factors. To study these factors in their interaction corpora with coreference annotation are needed. Recently, the task of creating such corpora not only for written texts, but also for different genres of spoken discourse has become topical.

One of the aims for coreference annotation in spoken discourse within the NLP paradigm is the frequency distribution of basic types of referring expressions

(full NPs vs. underspecified expressions such as pronouns and zeros) and the distribution of different features used by coreference resolution systems in spoken discourse in comparison to written texts. For this purpose, spontaneous speech undergoes a kind of normalization when different types of disfluencies are removed from the texts submitted to annotation process (for speech corpus normalization see (Fitzgerald and Jelinek, 2008; Hajič et al., 2008). Various kinds of disfluencies are annotated and studied separately at a separate level of annotation (Heeman et al., 2006).

Our study is based on the material of Pear story film retellings (Chafe, 1980) by healthy speakers of Russian and people with aphasia (PWA)—a language pathology resulting from damage to the language-dominant hemisphere of the brain. Each aspect of this topic has been well researched before. In the area of automatic text processing the task of developing corpora that include coreference annotation has been important for several decades (see for example the manual for coreference annotation (Chinchor and Robinson, 1997; Hirschman et al., 1997). A significant number of studies focus on different parameters and mechanisms involved in referential choice (see Fedorova, 2014; Kibrik, 2011 inter alia). The problem of Russian spoken discourse transcription, annotation, and analysis is covered by Kibrik and Podlesskaya (2009). The pear film has been used for four decades as elicitation stimulus for collection and analysis of narratives in a number of typologically different languages (Chafe, 1980; Erbaugh, 1990; Fedorova, 2014). Pear film retellings by English speakers are a part of a corpus with coreference annotation—ARRAU (Poesio and Artstein, 2008).

Unlike written discourse, spoken discourse has certain distinct features (Biber et al., 1999; Kibrik, 2009) such as hesitation pauses, self-corrections, discourse markers, and markers of word-finding difficulties (Bergelson et al., 2015; Podlesskaya and Kibrik, 2007; Shriberg, 1994). These disfluencies can affect the process of referent naming or the assessment of its prominence (extra referent mentioning attracts more attention to it and thus influences the referent prominence assessment). In clinical linguistics domain, analysis of reference in speech pathologies is not a common topic for research. The studies focus on finding differences between brain-damaged and healthy groups in the frequencies of basic classes of referential devices (full NP/anaphoric pronoun/zero pronoun) (Peng, 1992; Romanova, 2010), or on pronouns as means of establishing cohesion and coherence (Davis and Coelho, 2004). There are even more cases of disfluencies in the speech of the brain-damaged populations. The focus of this study as compared to the above mentioned is on the parameters that must be accounted for in coreferential chains annotation under the following condition: our data is comprised of the text retellings, not spontaneous production, including retellings by the brain-damaged individuals. We suggest some features to be employed for registering differences in written, spoken and clinical discourse.

The aim of this work is to describe issues that arise when the annotation scheme designed for written texts is adapted for spoken discourse analysis. In particular, we analyze the specific features of referring expressions in spoken discourse, including possible disfluencies and errors related to the referential choice.

2. Method and material

2.1. Participants and procedure

As mentioned above, our study is conducted on narratives from Russian CliPS (Clinical Pear Stories) corpus which contains Pear film (Chafe, 1980) retellings by people with aphasia (PWA) and neurologically healthy adults. The recorded narratives were transcribed and annotated with attention to speech failures and disfluencies in ELAN[2]. The narrative recording procedure, information about the speakers, and annotation scheme is described in (Khudyakova et al., 2016).

2.2. Coreference subcorpus

For the current study 11 texts by healthy speakers (norm) were chosen for developing annotation principles, and those principles were then applied to 9 texts by PWA. We have chosen texts by people with acoustic-mnestic and efferent motor aphasia (for description of aphasia types see, for example Akhutina, 2015; Luria and Hutton, 1977). Although PWA have deficits on micro-linguistic level, the narrative structure and coreference relations can be established (Marini, 2012). Lexical transcripts (with no annotation for pauses) of the texts were run through automatic lemmatizer and morphological analyzer[3]. The general statistics is shown in Table 1.

Table 1. The general statistics for the experimental corpus

	Healthy speakers	**PWA**
Number of texts	11	9
Min length in tokens	106	233
Max length in tokens	391	419
range	285	186
median	299	302
total	3,324	2,934

2.3. Annotation tool

As a starting point we have chosen to use the annotation scheme and annotation tool of RuCor corpus that was created for RU-EVAL forum on automated anaphora and coreference resolution (Toldova et al., 2014; http://ant0.maimbava.net/). Figure 1 demonstrates a fragment of the coreference annotation tool.

---

[2] https://tla.mpi.nl/tools/tla-tools/elan/

[3] We used Treetagger and lemmatizer for Russian http://corpus.leeds.ac.uk/mocky/

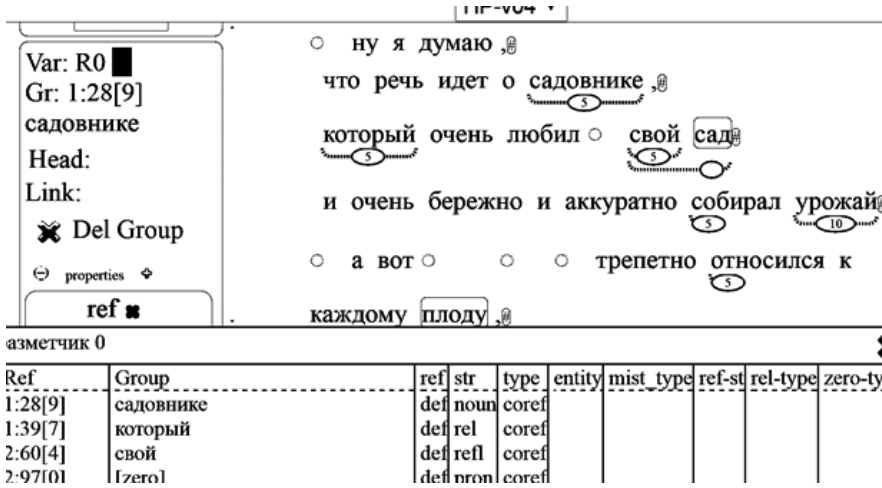


Figure 1. Annotation of coreference relations in the coreference annotation tool

When a coreference relation is established between two NPs, the same coreference chain ID is assigned to both NPs (see NP *садовник* '*a gardener*', *свой* '*his own*' and zero pronoun with the verb *берет собирал* 'collected'; they all have index 5 on the arc). The tool allows manual annotation of NP's head and embedded NPs, as well as assigning values for different NP features (the assigned values can be seen in the table on Figure 1). In case of ambiguity, it is possible to link NP to several coreference chains (cf. annotation scheme for "Pear stories" in Poesio and Artstein, 2008).

3. Annotation principles adaptation

To our knowledge none of the papers on coreference chain annotation in spoken discourse (see for example Poesio and Artstein, 2008) discussed the problem of annotating self-corrections, false starts etc., despite the fact that experimental research on reference shows that speech failures affect production and comprehension processes. For example, filled hesitation pauses in case of referential conflict facilitate its resolution and the choice of a newer referent as antecedent (Arnold et al., 2003).

In our coreference in spoken discourse annotation scheme we tried to pay attention to these phenomena and annotate not only specific features of coreference chain annotation, but also speech failures related to production and choice of an appropriate referential expression.

3.1. Markable boundaries

Many international standards for annotation of written texts define a markable as the maximally full NP up to the nearest comma to the right (see for example

Krasavina and Chiarcos, 2007). However, spoken discourse has certain features that do not allow using this criterion.

Markable borders in case of sequential nomination of a referent ('renaming constructions')

There are cases where two NPs denoting the same entity participate in so-called "renaming" constructions, e.g. an alternative construction, as in (1). Such constructions are means to express overtly the mental process of referent naming, that is—seeking a proper referring expression, correcting or refining the chosen one, as in (2) (see Bergelson et al., 2015 for more detailed discussion).

(1) [*мальчик*] [*или парень*]
'a boy [*or a guy*]'

(2) [*груши*], [*или, как там их, грушины*]
'pears, [*or how do you call them, pear-things*]'.

Such cases pose a problem for classical approach to coreference annotation. They are not coreferential expressions in the proper sense of the word, but they also can influence referential choice by being a factor of additional referent activation (see Givon's notion of topicality in Givón, 1983).

We decided to define as separate markables all NPs that name one referent in one point of discourse, and to establish a special type of relation between them rather than include them in the chain as coreferent NPs. The connectives, discourse markers and parenthetical words are also included into markable.

Markable boundaries under non-standard syntactic environment

In retellings various discourse markers can be embedded in or adjacent to NPs though their standard syntactic position is the sentential modifier. These markers can be attributed to the degree of confidence of referential choice or interaction components of the discourse:

(3) *поглощающие по внешнему виду его груши* (HP-v02)
'Consuming as it looks his pears'

Markables in case of postpositional adjectival modifiers: apposition vs. an entire NP

Adjectival phrases in postposition to the head in a referring expression pose a separate problem. When annotating written texts, one would normally use a principle of the 'left' punctuation border. In retellings speakers not only verbalize the procedure of choosing the most appropriate referential device, but also the procedure of 'attributive description choice' for the referent. That is why adjectival phrases in postposition are quite common in oral narratives:

(4) … *и прошли как раз мимо* [*хозяина груши этой большой*]
Lit. 'And passed by the owner [*of the pear tree this big*]' (c.f. this big pear tree)

In spoken discourse annotation the "left punctuation border" criterion cannot be applied. Postpositive adjectival phrases (such as (4) can be interpreted as parcellation, however, when descriptions in postposition appear in written texts, they are usually characterized by a certain syntactic structure (see Ljutikova, 2015). Our decision was to place such descriptions into the same markable unless any specific signs of border (e.g. declining intonation and long pause) are present. However, postposition of the adjectival phrase is reflected in the NP morphosyntactic structure parameter.

Split NPs

In spoken discourse split NPs are more prevalent, as well as non-projectiveness and non-canonical word order, as in (5):

(5) *А куда* [*корзина-то*]$_{gr1}$ *делась* [*одна*]$_{gr1}$ (HP-v03)
And somewhere [*the basket*]$_{gr1}$ *has gone4* [*one*]$_{gr1}$

In (5) the numeral *одна* '*one*' is in postposition to its head (cf. *одна корзина* '*one basket*') and is separated from it by the verb.

3.2. Entity types

The nature of texts in the corpus, which are retellings of the same film makes it somewhat easier to analyze coreference, because the characters of the story are the same for all narratives. The speaker can pick an inadequate referential device, use a deictic device *вот этот* '*this* one' or anaphoric pronoun *он* '*he*' without introducing the referent, but in this case the annotator would still be able to refer it to the appropriate NP based on the context and the annotator's knowledge of the story. Besides, in the film a number of characters and objects belong to the same ontology class and can be referred to by the same name, which allows for the referential conflict to appear not only in case of anaphoric pronouns, but also full NPs.

To differentiate entities in the narrations we included the special labels such as 'man', 'baskets_man', 'boy', 'three_boys' etc. This type of annotation allows us to compare the referential expressions used by different speakers for the same entities, and different expressions used for the entities from the same ontological class.

3.3. Morphosyntactic types of NPs

Our investigation of the NP morphosyntactic properties for markables in the corpus has shown that there are some special cases that are characteristic of oral narration both by PWA as well as healthy speakers.

Occurrences of NPs with demonstratives as the introductory NPs, e.g. *И вот вдруг приехал этот мальчик* '*And suddenly this boy came*') are not common in written texts. Moreover, the NPs with demonstratives are rare in Russian news texts (Nedoluzhko et al., 2015).

Another peculiarity that was revealed for referential devices in spoken discourse is that a parenthetical word can serve as a prenominal modifier, as in (6) and (7):

(6) *залез на, видимо, груши* (HP-v01)
'Got onto, obviously, pear'

(7) *ну подъехал его, наверное, его сын младший*
'And came his, maybe, his son younger'

As it was mentioned earlier, an important parameter is the place of the modifier relative to the NP head *груши спелые 'pears ripe'* vs. *спелые груши 'ripe pears'*.

(8) *… в* [*такой шапке летней*]
'... in [*such a hat summer-y*]

The following morphosyntactic types taken from RuCor annotation scheme are retained: NPs with demonstratives, NPs with other modifiers (adjectives, numerals, indefinite pronouns); bare nouns, anaphoric pronouns (both 3rd person pronouns and reflexives), relative pronouns and zero pronouns (pro). In order to check some specific features of NPs used in oral retellings we use the more detailed classification of NPs: the type of pronoun or numeral is taken into consideration, the type of modifiers, as well as the type of word order.

3.4. Types of zeroes

Russian is a so called pro-drop language, that is finite clauses with no overt subject are possible as well as 'omitted' anaphoric pronouns in some other positions. While zero subjects are very rare in news texts, zero subjects chaining is a standard strategy for spoken discourse. Number of zero pronouns can be an important parameter in the analysis of pathological discourse compared to healthy discourse. Each predicate belongs to an elementary discourse unit (EDU), and, thus, we restore zero subjects for all the verb forms with no overt subjects.

We added distinction of different types of zero pronouns into the annotation scheme: syntactically motivated zeroes, conjunction zeroes, subject zeroes in separate clauses.

3.5. Link types in chains: annotation of non-coreference relations between NPs

Naming relations: renaming and speech disfluences

There are cases in corpus when NPs do not refer directly to an entity rather they denote the process of referent's naming or they represent speech disfluencies that can affect the degree of entity activation (for the factors influencing referemt's a\ctivation

see Kibrik A. A. 2011): false-starts, repetitions, self-corrections (10), (11); name elaboration (12) and alternative naming (9):

(9) *Плоды* [*груши*] [***или*** *авокадо*] (HP-v02)
'Fruit of pears or avocadoes'

(10) *Яблоки,* [***точнее***, *груши*]
'Apples, no, pears'

(11) *И поставили корзину*$_i$ *на витрину*$_j$, *ведро*$_i$ *на велосипед*$_j$ (AP-v01)
'And put the basket on the window bucket on the bike'

(12) *Ребята* ***ну*** *друзья его из деревни* (AP-v07)
'Guys well friends his from the village'

For some occurrences of referential expressions, NP elaboration is hard to distinguish from NPs in postposition (12).

Self-corrections may apply to a NP due to the wrong choice of its referential status:

(13) *Эту грушу каждую грушу вытирал*
'This pear (wrong referential expression) *every pear*

Special cases of apposition links

A special type of apposition is used in spoken discourse (it does not occur in written texts). That is an anaphoric pronoun followed by full NP (14). In spoken discourse they reflect monitoring the process of hearer's referential expression interpretation by the speaker.

(14) *И забрал* [*ее*], [*корзину*]
And he took [*it*], [*the basket*]

The following basic annotation features are used for written texts: link type (coreference, apposition, predicative (Toldova et al. 2014)). Besides these link types, we have introduced additional values of the feature 'Link Type', namely 'repetition', 'self-correction', false start', 'alternative nomination' and some others.

3.6. Error types

Spoken discourse demonstrates various errors in naming and in choice of NPs, that's why we introduced a specific parameter to capture these errors. A detailed typology of these errors requires additional research. As mentioned in (Bergelson et al. 2015) referential errors are caused by different mechanisms of speech production. In our initial annotation we pay attention to only basic type of errors:

(a) morphological errors—the speaker chose the wrong number or person agreement marker, or a wrong case marker (15):

(15) *То ли не все положили, остались у него* (у них) *груши в руках*
'As if not all were placed, pears remained in his (in their) *hands'*

(b) wrong lexical choice (semantic paraphasia) (*мешок* '*bag*' instead of *корзина* '*basket*')
(c) wrong choice of referential expression, like *эти мальчики* '*these boys*' instead of *три мальчика* '*three boys*' when introducing the referent.

(16) *Эту грушу каждую грушу вытирал*
'He wiped this pear, every pear'

3.7. Summary statistics

Deviations of the coreference annotation scheme for spoken discourse from that for the written texts reflect specific features of coreference in the former—both the speech of healthy people as well as the aphasic speech.

The comparison of basic morphosyntactic types distribution for written vs. spoken discourse is given in Table 2. The figures for written texts are taken from (Nedoluzhko et al., 2015) where the written text corpus consists of 16 short news texts on political and economic topics (the average length is 30 sentences). The figures are given for the markables including appositions and excluding various kinds of renaming and disfluencies (e.g. repetitions and false starts).

Table 2. The distribution of morphosyntactic types of referential devices in written texts, retellings by neurologically healthy people and by PWA

NP morphosyntactic type		Written texts		Pear Stories			
				Healthy speakers		PWA	
anaphoric and reflexive pronouns	subject position	39	3.8%	148	15%	138	14%
	non-subject position	95	9.3%	145	14%	112	11%
relative		42	4.1%	19	2%	21	2%
zero (pro)		13	1.3%	199	20%	196	20%
bare noun		164	16.0%	338	33%	338	33.1%
NP with a demonstrative		20	1.9%	49	5%	38	4%
Other NPs		652	63.6%	163	18%	131	13%
TOTAL		**1,025**	**100%**	**1,061**	**100%**	**974**	**100%**

As demonstrated in table 2. the reduced referential expressions (pronouns and zero pronouns) are much more frequent in retellings than in written news texts. There is a great difference in the anaphoric pronouns distribution for news texts vs. retellings. However, the difference on pronoun frequency between healthy speakers' texts and PWA texts is not so substantial. The more striking contrast is in zero anaphora distribution (the frequency is 15 times greater in in retellings than in written texts). It is also worth mentioning that the distribution of demonstratives is significantly lower in written texts as compared to spoken discourse.

As for different types of disfluencies, they make approximately 8% of all markables for the healthy speakers and 10% for the PWA.

4. Conclusions

Disfluencies represent one of the most eye-catching features of spoken discourse. They mark the process of speech production directly in the resulting text. Often when performing coreference chains annotation for spoken discourse the text is 'purified' from disfluencies and interaction markers. It means that two objects—a 'normalized' text and various disfluencies are studied as separate systems. At the same time presence of disfluencies in the text has impact on the interpretation of other text elements and also on the speaker's verbalization choices. While adapting the initial coreference annotation scheme we came to a conclusion that besides the referential ambiguity, which is normally taken into account in spoken discourse analysis, and basic taxonomy of the referential devices (full NP vs. anaphoric pronoun vs. anaphoric zero) we need to include there both disfluencies and interactional markers.

Thus, we suggest an approach to the coreference relations annotation of spoken discourse that integrates various phenomena. Those are coreferential narrative chains, disfluencies (like changing the name of the referent) and interactional elements (for instance, speakers' assessment of the correctness of their choice of nomination).

References

1. *Akhutina, T.* (2015). Luria's classification of aphasias and its theoretical basis. Aphasiology, 1–20. doi:10.1080/02687038.2015.1070950.
2. *Arnold, J. E., Fagnano, M., and Tanenhaus, M. K.* (2003). Disfluencies signal theee, um, new information. in Journal of Psycholinguistic Research, 25–36.
3. *Bergelson, M. B., Akinina, Y. S., Dragoy, O. V., Iskra, E. V., and Khudyakova, M. V.* (2015). Markers of word production difficulties in normal and clinical discourse production: continuity of norm in language and discourse [Zatrudnenija pri porožděnii slov v diskurse i ix formal'nые markery: norma i patologija, ili o nediskretnosti normy v jazyke. Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference Dialogue 14, 41–51.
4. *Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E.* (1999). The Longman grmmar of spoken and written English.

5. *Chafe, W.* (1980). The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production. , ed. W. Chafe Norwood, New Jersey: Ablex.
6. *Chinchor, N., and Robinson, P.* (1997). MUC-7 named entity task definition. in Proceedings of the 7th Conference on Message Understanding, 29.
7. *Davis, G. A., and Coelho, C. A.* (2004). Referential cohesion and logical coherence of narration after closed head injury. Brain and Language 89, 508–523. doi:10.1016/j.bandl.2004.01.003.
8. *Erbaugh, M. S.* (1990). Mandarin Oral Narratives Compared with English: The Pear/Guava Stories. Journal of the Chinese Language Teachers Association 25, 21–42.
9. *Fedorova, O. V* (2014). Experimental discourse analysis [Eksperimental'nyj analiz diskursa]. Languages of Slavonic Culture.
10. *Fitzgerald, E., and Jelinek, F.* (2008). Linguistic resources for reconstructing spontaneous speech text. in LREC Proceedings (Marrakech, Morocco), 1–8.
11. *Givón, T.* (1983). "Topic Continuity in Discourse: An Introduction," in Topic Continuity in Discourse: A Quantitative Cross-language Study, ed. T. Givón (John Benjamins), 3–41.
12. *Hajič, J., Cinková, S., Mikulová, M., Pajas, P., Ptáček, J., Toman, J., et al.* (2008). An annotated resource for speech reconstruction. in 2008 IEEE Workshop on Spoken Language Technology, SLT 2008—Proceedings, 93–96.
13. *Heeman, P. a, McMillin, A., and Yaruss, J. S.* (2006). An annotation scheme for complex disfluencies. INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing 3, 1081–1084.
14. *Hirschman, L., Robinson, P., Burger, J., and Vilain, M.* (1997). Automating Coreference : The Role of Annotated Training Data. in Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloqium.
15. *Khudyakova, M. V., Bergelson, M. B., Akinina, Y. S., Iskra, E. V., Toldova, S., and Dragoy, O. V.* (2016). Russian CliPS: a Corpus of Narratives by Brain-Damaged Individuals. in LREC Proceedings (Portoroz, Slovenia).
16. *Kibrik, A. A.* (2009). Modus, genre and other parameters of discourse classification [Modus, zhanr i drugie parametry klasifikatsii diskursov]. Topics in the study of language [Voprosy Jazykoznanija] 2, 3–21.
17. *Kibrik, A. A.* (2011). Reference in discourse. Oxford University Press.
18. *Kibrik, A. A., and Podlesskaya, V. I. eds.* (2009). Night Dream Stories: A corpus study of spoken Russian discourse [Rasskazy o snovidenijah: korpusnoe issledovanie ustnogo russkogo diskursa]. Moscow: Languages of Slavonic Culture.
19. *Krasavina, O. N., and Chiarcos, C.* (2007). PoCoS: Potsdam coreference scheme. in Proceedings of the Linguistic Annotation Workshop (Association for Computational Linguistics), 156–163.
20. *Ljutikova, E.* (2015). Agreement, features and structure of NP in Russian [Soglasovanie, priznaki i struktura imennoj gruppy v russkom jazyke]. Russian Language and Linguistic Theory [Russkij yazyk v nauchnom osveshchenii] 2.
21. *Luria, A. R., and Hutton, J. T.* (1977). A modern assessment of the basic forms of aphasia. Brain and Language 4, 129–151.

22. *Marini, A.* (2012). Characteristics of narrative discourse processing after damage to the right hemisphere. Seminars in Speech and Language 33, 68–78. doi:10.1055/s-0031-1301164.
23. *Nedoluzhko, A., Toldova, S., and Novák, V.* (2015). Coreference Chains in Czech, English and Russian: Preliminary Findings. Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" 14, 456–469.
24. *Peng, V. M.* (1992). The usage of reference items in aphasic and normal conversations. Journal of Neurolinguistics 7, 295–307. doi:10.1016/0911-6044(92)90020-W.
25. *Podlesskaya, V. I., and Kibrik, A. A.* (2007). Self-corrections of the speaker and other types of speech failures as object of annotation in spoken corpora [Samoispravlenija govorjaščego i drugie tipы rečevыx sboev kak ob"ekt annotirovanija v korpusax ustnoj reči]. Science-Technical Information [Naučno-texničeskaja informacija] 2, 2–23.
26. *Poesio, M., and Artstein, R.* (2008). Anaphoric annotation in the ARRAU corpus. in Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08) (Marrakech, Morocco).
27. *Romanova, A.* (2010). Referential Choice: Distribution of Subject Types in Russian Aphasic Speech. The realization of L*+ H pitch accent in Greek, 139–147.
28. *Shriberg, E. E.* (1994). Preliminaries to a theory of speech disfluencies.
29. *Toldova, S., Roytberg, A., Ladygina, A. A., Vasilyeva, M. D., Azerkovich, I. L., Kurzukov, M., et al.* (2014). RU-EVAL-2014 : Evaluating Anaphora and Coreference Resolution for Russian. Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference Dialogue 14, 1–14.

MULTIPLE FEATURES FOR MULTIWORD EXTRACTION: A LEARNING-TO-RANK APPROACH

Tutubalina E. V. (tutubalinaev@gmail.com)
Kazan Federal University, Kazan, Russia

Braslavski P. I. (pbras@yandex.ru)
Ural Federal University, Yekaterinburg, Russia

This paper describes the extraction of multiword expressions (MWEs) from corpora for inclusion in a large online lexical resource for Russian. The novelty of the proposed approach is twofold: 1) we use two corpora—the Russian National Corpus and Russian Wikipedia—in parallel and 2) employ an extended set of features based on both data sources. To combine syntactic and statistical features derived from two corpora, we experiment with several learning-to-rank (LETOR) methods that have been proven to be highly effective in information retrieval (IR) scenarios. We make use of bigrams from existing dictionaries for learning, which leads to very sparing manual annotation efforts. Evaluation shows that machine-learned rankings with rich features significantly outperform traditional corpus-based association measures and their combinations. Analysis of resulting lists supports the claim that multiple features and diverse data sources improve the quality of extracted MWEs. The proposed method is language-independent.

Key words: multiword expressions (MWEs), collocations, lexical acquisition, learning-to-rank methods (LETOR), thesaurus, Russian language

ИЗВЛЕЧЕНИЕ МНОГОСЛОВНЫХ ВЫРАЖЕНИЙ НА ОСНОВЕ МНОЖЕСТВЕННЫХ ПРИЗНАКОВ И МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ РАНЖИРОВАНИЮ

Тутубалина Е. В. (tutubalinaev@gmail.com)
Казанский федеральный университет, Казань, Россия

Браславский П. И. (pbras@yandex.ru)
Уральский федеральный университет, Екатеринбург, Россия

Ключевые слова: многословные выражения, устойчивые словосочетания, машинное обучение ранжированию, тезаурус, русский язык

1. Introduction

Multiword expressions (MWEs) are heavily underrepresented in existing Russian lexical resources. We encountered the problem of MWE extraction within a project aimed at creating a new wordnet for Russian. The study described in the paper deals with nominal bigrams—the most common MWE type. Since the pioneering work by Church and Hanks (1989) the problem of MWE extraction has been studied in depth, and various statistical association measures (AMs) have been proposed. Despite the task seems to be solved, larger datasets and advanced statistical methods available nowadays offer opportunities for a more efficient solution.

The proposed approach includes three components: 1) we use two different corpora—the Russian National Corpus (364M tokens) and Russian Wikipedia (1.2M articles, 318M tokens)—in parallel; 2) MWE candidates are described with a rich set of features (various corpus-based statistics, link-based Wikipedia features, phrase structure, Web statistics, etc.); 3) we formulate the MWE ranking task in terms of multiple 'queries' and 'documents' and apply learning-to-rank (LETOR) algorithms that showed good results in information retrieval (IR) scenario. Our approach deals with different kinds of MWEs—collocations, idioms, set phrases, etc. (see classification in Baldwin and Kim (2010))—in a uniform way. We took several thousands of nominal bigrams from existing Russian dictionaries and manually labeled them as positive and negative examples. This routine allowed us to minimize manual labeling efforts and is more advantageous than such alternatives as labeling output of an automatic method, which can potentially introduce bias towards presented results, or asking an expert to produce a list of good and bad examples from scratch, which is very labor-intensive. Using limited training data, we were able to rank the whole set of candidate MWEs extracted from both corpora and cut them off at desired level (our estimate of the target number of MWEs for the wordnet under development is around 40K). Evaluation showed that proposed approach outperformed existing AMs, as well as classification-based methods. Manual probes proved that high-ranked MWEs are good enough to be included in the resource with minimal manual intervention. The method is language-independent—it relies only on the availability of a large corpus, Wikipedia, and a part-of-speech (POS) tagger. Furthermore, the method is highly flexible and can be applied to other MWE types.

2. Related Work

There is a large body of literature on extraction of multiwords, collocations, and keyphrases; Hasan (2014) and Ramisch (2015) provide an extensive overview of the field. Three groups of approaches related to our work can be distinguished: (i) methods based on purely statistical AMs; (ii) machine-learned classification; (iii) Wikipedia-based approaches to terminology extraction.

Traditional approaches rank a list of MWEs according to their co-occurrence frequencies or statistical AMs (Evert and Krenn, 2005; Pecina and Schlesinger, 2006). Krenn and Evert (2001) evaluated Mutual Information (MI), Dice coefficient, Student's t-score and log-likelihood ratio for adjective-noun pairs. Pecina and Schlesinger

(2006) evaluated 82 measures on a Czech corpus. Some studies suggested different strategies for handling low-frequency and high-frequency items (Evert and Krenn, 2001; Evert and Krenn, 2005; Bouma, 2009). Wermter and Hahn (2006) showed that the most advanced AMs perform similarly to raw frequency.

State-of-the-art studies consider the MWE extraction task as a classification problem (Pecina and Schlesinger, 2006; Fothergill and Baldwin, 2011; Karan et al., 2012; Ramisch, 2015). Pecina and Schlesinger (2006), Ramisch et al. (2010) and Karan et al. (2012) employed support vector machines (SVM) with frequency counts, traditional AMs, and POS patterns as features. These supervised approaches are different from ours in that Karan et al. (2012) and Ramisch (2015) created a training set consisting of positive and negative MWE examples, while Fazly and Stevenson (2007) and Fothergill and Baldwin (2011) assigned MWE categories. Feature-rich ranking of keyphrases extracted from a document is close to our approach (Jiang et al., 2009). However, extracting keyphrases from a document exploits quite a different set of document-level features such as position of the first occurrence, document field (e.g. title, section heading, anchor text), and text highlighting (e.g. boldface). Document-level keyphrase extraction task differs from our setting in that the same word sequence occurring in different documents can be a good keyphrase in one case, but not suitable in other cases.

Many studies explored Wikipedia as an external knowledge resource for terminology extraction (Hartmann et al., 2012; Vivaldi et. al., 2012) and keyphrase extraction (Medelyan et al., 2009). Medelyan et al. (2009) used a machine learning approach with Wikipedia-based semantic features to determine whether the document can be annotated with a given keyphrase. Hartmann (2012) considered n-grams that appeared in Wikipedia titles and anchor text as candidates for subsequent ranking by AMs. (Vivaldi et al., 2012) used Wikipedia categories to validate term candidates extracted from scientific texts.

Due to limited space we do not survey a large body of literature on learning to rank and feature selection for IR; Liu (2009) gives a nice overview of approaches and methods. In our work we follow the feature selection approach proposed by Geng et al. (2007) that combines two scores: importance of individual features and similarity between features.

3. Data

In our study, we use two corpora—Russian National Corpus[1] (RNC) and Russian Wikipedia[2]. RNC has genre subdivisions—scientific texts, classical literature, legal and official documents, religious texts, children's literature, nonfiction, news, etc.—that we use for feature calculation. We treat Wikipedia both as a "plain text corpus" to calculate MWE statistics and as semi-structured data: we make use of Wikipedia links, redirects, categories, and page titles. Lemmatization and POS-tagging is performed with *mystem* library[3].

---

[1] http://ruscorpora.ru/en

[2] http://ru.wikipedia.org

[3] https://tech.yandex.ru/mystem

We consider all bigrams conforming to one of six morpho-syntactic patterns—*Adjective + Noun, Noun + Adjective, Participle + Noun, Noun + Participle, Noun + Noun (genitive)*, and *Noun + Noun (instrumental)*—as candidate MWEs. Moreover, a candidate MWE must occur at least ten times in the RNC or to be a Wikipedia title.

We also collected nominal bigram entries from three dictionaries: Wiktionary[4] (3,155), Small Academic Dictionary (2,955), and Ushakov's Dictionary (2,506), which resulted in 7,751 unique bigrams in total. Manual inspection revealed that the list contained many archaisms (e.g. *книга живота—book of life*, *духовное брашно—spiritual repast*), narrowly used metaphorical expressions (e.g., *деревянный макинтош—coffin* (literally—*wooden mackintosh*), *белый друг—toilet bowl* (*white friend*)), joking expressions (e.g., *губозакаточная машинка—lip-rolling machine*), as well as named entities (e.g. *Амурская область—Amur Region*). The list underwent manual labeling by two lexicographers. Lexicographers labeled MWEs as positive (suitable for a general-purpose thesaurus) and negative (otherwise). Manual labeling resulted in an approximately equal number of positive and negative examples. Table 1 summarizes the data used in the study.

Most advanced LETOR algorithms (so-called pair-wise and list-wise methods, see (Liu, 2009)) optimize ranking in the context of individual queries and respective result lists in contrast to earlier point-wise approaches that model relevance as global regression or classification task. In order to apply modern LETOR algorithms to the MWE extraction task, we represent the data as a set of "queries" and "documents". Our hypothesis is that 'divide and conquer' approach helps deal with MWEs of different types and frequency ranges in a unified way in the learning phase. For "queries", we took 5,871 unique words from labeled examples to create individual lists of MWE candidates ("documents") containing the "query" (see Table 2). 56.5% of all candidates were included at least in one list. We randomly sampled 80% of the 'queries' for training and held out 20% for testing.

Table 1. Candidate MWEs and labeled data (overlapping bigrams have at least one common word)

# of positive examples	3,981
# of negative examples	3,770
# of unique words in labeled examples	5,871
# of positive examples in the test set	1,322
# of candidate MWEs from the RNC	190,416
# of candidate MWEs from Wikipedia	157,748
# of unique candidate MWEs from both corpora	329,866
# of candidate MWEs overlapping with labeled set (RNC)	82,456
# of candidate MWEs overlapping with labeled set (Wiki)	117,837
# of unique candidate MWEs overlapping with labeled set	188,441

[4] http://ru.wiktionary.org

Table 2. 'Queries' (single words from labeled bigrams) and 'documents' (overlapping candidate MWEs); positive examples are underlined

word ('query')	overlapping bigrams ('documents')
неправильный	неправильная установка (wrong installation), неправильная постановка (wrong statement), неправильная музыка (wrong music), неправильная галактика (wrong galaxy), неправильная переменная (wrong variable), неправильная дробь (improper fraction)
струна	слабая струна (weak string), натянутая струна (tense string), гетеротическая струна (heterotic string), бозонная струна (bosonic string), квантовая струна (quantum string), золотая струна (gold string), космическая струна (cosmic string), спинная струна (notochord)
корова	белая корова (white cow), старая корова (old cow), черная корова (black cow), синяя корова (blue cow), священная корова (sacred cow), дойная корова (milk cow), морская корова (sea cow)
вещество	специальное вещество (special substance), обычное вещество (usual substance), рабочее вещество (working substance), солнечное вещество (solar substance), сухое вещество (solid), белое вещество (white substance), мягкое вещество (soft substance), полярное вещество (polar substance), компактное вещество (compact substance), радиоактивное вещество (radioactive material), живое вещество (live substance), лекарственное вещество (medicinal substance), действующее вещество (active ingredient), вредное вещество (harmful substance), серое вещество (gray substance), простое вещество (simple substance), органическое вещество (organic), химическое вещество (chemical agent)

4. Methods

To apply a ranking algorithm to the data we have to present each MWE candidate as a feature vector. Note that, in the IR scenario a vector represents a query-document pair, i.e. there are features depending on the query, document, or both. In our case, all features describe an individual MWE independently from the "query", which allows us to apply the obtained ranking function later to the global set of candidates (hundreds of thousands items). The feature set used in the study (42 features in total) is described below.

RNC features (14): RNC global frequency, ten frequencies in genre subcorpora (reflects specificity of the MWE), first and second words' frequencies, the presence of the candidate in the corpus.

Wikipedia-based features (20) included: Wikipedia frequency, the presence of a redirect with the given MWE, match with a Wikipedia title, the number of in- and out-links, the number of categories assigned to the page, the presence of an infobox,

11 binary features corresponding to the infobox type[5], and capitalization (the latter three features aimed at capturing named entities).

Structural features (7) included six binary features corresponding to the above mentioned extraction patterns plus bigram length in characters (indirectly reflects the bigram specificity).

Web document frequency (1) refers to the number of documents returned to MWE as a phrase query by a search engine (SE) through an API[6].

We used three algorithms implemented in the RankLib library[7] to obtain MWE rankings: MART (Friedman, 2001), RankBoost (Freund et al., 2003), and LambdaMART (Wu et al., 2007) with default parameters. To improve efficiency of the training, we applied a feature selection (Geng et al., 2007). We held out 20% of the training set as validation set to optimize the number of features. First, according to the method, we computed importance of each feature using *mean reciprocal rank* (MRR). We measured similarity between features with Kendall's τ for pairs of corresponding rankings. Second, we maximized the sum of the importance scores of individual features and minimized the total similarity score between the features using a greedy search algorithm. Finally, five groups of features with the best results on the validation set were used to evaluate LETOR models on the test set.

5. Evaluation

We evaluated multiple intermediate rankings with artificial queries using two measures: 1) *mean reciprocal rank* (MRR) and 2) *bpref*, an evaluation measure suited for incomplete judgments (Buckley and Voorhees, 2004). MRR is an average of inverse ranks of the first positive example in each 'query'; while *bpref* accounts for inversions—cases, when 'relevant' items are ranked lower than 'non-relevant' ones. Both measures were averaged over 1,449 lists in the test set.

We compared our approach to state-of-the-art collocation extraction methods based mainly on frequency (Pecina and Schlesinger, 2006; Ramisch et al., 2010; Karan et al., 2012). In particular, we implemented the best-performing method for keyphrase extraction (Jiang et al., 2009) based on SVM-rank[8] algorithm and following features: POS patterns, MWE frequency, and 20 AMs calculated using UCS toolkit[9] on (i) RNC, (ii) Wikipedia, (iii) both corpora. We also implemented AMs (t-score, log-likelihood, and MI) as baselines. Evaluation results are presented in Table 3.

---

[5] http://en.wikipedia.org/wiki/Wikipedia:List_of_infoboxes

[6] https://xml.yandex.ru/

[7] http://sourceforge.net/p/lemur/wiki/RankLib

[8] http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

[9] http://www.collocations.de/software.html

Table 3. MRR and *bpref* measures computed on the test set

Ranking method	MRR	*bpref*
MI	0.440	0.353
t-score	0.615	0.321
log-likelihood	0.620	0.353
Wikipedia frequency	0.625	0.467
RNC frequency	0.624	0.328
SVM-rank (RNC)	0.644	0.550
SVM-rank (Wikipedia)	0.609	0.492
SVM-rank (RNC+Wikipedia)	0.635	0.483
MART	0.639	0.545
MART + feature selection	0.639	0.480
LambdaMART	0.679	0.742
LambdaMART + feature selection	0.684	0.546
RankBoost	0.739	0.742
RankBoost + feature selection	**0.758**	**0.825**

As the results show, LambdaMART and RankBoost scored best compared to MART, SVM-Rank and AMs. SVM-rank and MART scores are comparable. Impact of feature selection is mixed: it improved both MRR and *bpref* for RankBoost, but degraded LambdaMART and MART *bpref* scores. Best LambdaMART results were obtained with all features except for the Wikipedia title feature and four Wikipedia infobox features. RankBoost scored best using the Wikipedia title feature, number of categories, and presence of candidate in the corpus. Table 4 illustrates the contribution of different feature groups to the overall performance. The results support our initial hypothesis that multiple data sources improve results.

Table 4: MRR for highest and lowest positive items ranked with LambdaMART: contribution of different feature groups

	MRR highest	MRR lowest
all features	0.679	0.598
w/o RNC-based features	0.565	0.497
w/o Wikipedia-based features	0.609	0.543
w/o structural features	0.671	0.592
w/o results from the search engine	0.678	0.602

Top-40K lists ranked by LambdaMART, SVM-Rank, and RNC-based frequency contain 634, 472, and 452 positive examples (out of 990 'relevant' MWEs in the initial global list), respectively. In the top-40K MWEs ranked by LambdaMART, 43% items occur in both corpora, 35% and 22% occur in Wikipedia or RNC only, respectively. This again illustrates the benefit of using two data sources in parallel. Figure 1 presents ROC curves for the top-40K candidate MWEs ranked by LambdaMART, SVM-Rank, and RNC-based frequency (note that the total number of true positives differs for these 40K-lists, see above). Table 5 shows MWEs at different levels of the global list ranked by LambdaMART.

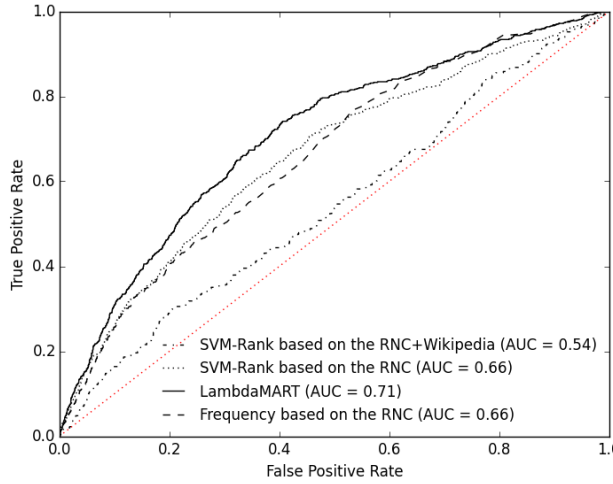


Fig. 1: ROC curves for four methods

Table 5: Examples of MWEs at different levels of the global ranking

Cut-off level = 100	**Cut-off level = 1,000**
земная кора (Earth's crust) программное обеспечение (software) основные фонды (basic assets) биологические науки (bioscience) общественное мнение (public opinion)	подсадная утка (decoy-duck) народный дух (national character) разговорная речь (spoken language) публичная библиотека (public library) братская могила (mass grave)
Cut-off level = 2,500	**Cut-off level = 5,000**
диалектическая логика (dialectical logic) барионный заряд (baryon charge) врождённые идеи (innate idea) гонка вооружений (arms race) адский огонь (hellfire)	фразовое ударение (phrasal stress) блуждающие огни (will-o'-the-wisp) золотой телец (golden calf) циркуляция крови (blood motion) кольцевые гонки (circuit race)
Cut-off level = 10,000	**Cut-off level = 30,000**
грудная железа (breast gland) критическая теория (critical theory) чесменский бой (battle of Chesma) автоматический огонь (automatic fire) личное дворянство (personal nobility)	концептуальное искусство (conceptual art) институциональный инвестор (institutional investor) земские марки (zhemstvo stamps) агглютинативные языки (agglutinative language) ненасыщенный пар (unsaturated steam)
Cut-off level = 100,000	**Cut-off level = 150,000**
шлиховой анализ (panning) облеченный тон (invested tone) дардские народы (dardsky people) трамвайная археология (tram archeology) глухой удар (bump)	ноги прохожих (feet of passers-by) разделенный экран (divided screen) воркутинская улица (Vorkuta street) осетинская церковь (Ossetian church) старый базар (old market)

6. Conclusion

In this paper, we described an experiment on MWE extraction from corpora. The novelty of the approach lays in the use of two data sources in parallel, a rich set of features, and advanced learning-to-rank methods applied to the task. The proposed approach outperforms traditional association measures and state-of-the-art classification methods. The method is language-independent and employs limited training data. In the future, we plan to apply the method to the extraction of verbal MWEs.

Acknowledgments

Work on problem definition, survey of related work, experimental design, and feature engineering was carried out by Pavel Braslavski and supported by the Russian Foundation for Basic Research grant no. 14-37-50953. Work on method implementation and evaluation was carried out by Elena Tutubalina and supported by the Russian Science Foundation grant no. 15-11-10019.

References

1. *Buckley Chris and Voorhees Ellen M.* (2004), Retrieval evaluation with incomplete information, In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 25–32.
2. *Church Kenneth Ward and Hanks Patrick* (1990), Word association norms, mutual information, and lexicography, Computational linguistics, Vol. 16(1), pp. 22–29.
3. *Evert Stefan and Krenn Brigitte* (2005), Using small random samples for the manual evaluation of statistical association measures, Computer Speech & Language, 19(4), pp. 450–466.
4. *Evert Stefan and Krenn Brigitte* (2001), Methods for the qualitative evaluation of lexical association measures. In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, pp. 188–195.
5. *Fazly A. and Stevenson S.* (2007), Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures, In Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, Association for Computational Linguistics, pp. 9–16.
6. *Fothergill Richard and Baldwin Timothy* (2011), Fleshing it out: A supervised approach to mwe-token and mwe-type classification, In IJCNLP, pp. 911–919.
7. *Freund Yoav, Iyer Raj, Schapire Robert E, and Singer Yoram* (2003), An efficient boosting algorithm for combining preferences, The Journal of machine learning research, Vol. 4, pp. 933–969.
8. *Friedman Jerome H.* (2001), Greedy function approximation: a gradient boosting machine, Annals of statistics, pp. 1189–1232.

9. *Geng Xiubo, Liu Tie-Yan, Qin Tao, and Li Hang* (2007), Feature selection for ranking, In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 407–414.
10. *Hartmann Silvana, Szarvas György, and Gurevych Iryna* (2012), Mining multiword terms from Wikipedia, Semi-Automatic Ontology Development: Processes and Resources, pp. 226–258.
11. *Hasan Kazi Saidul and Ng Vincent* (2014), Automatic keyphrase extraction: A survey of the state of the art, Proceedings of the Association for Computational Linguistics (ACL), Baltimore, Maryland: Association for Computational Linguistics.
12. *Jiang Xin, Hu Yunhua, and Li Hang* (2009), A ranking approach to keyphrase extraction. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, ACM, pp. 756–757.
13. *Karan Mladen, Snajder Jan, and Basic Bojana Dalbelo* (2012), Evaluation of classification algorithms and features for collocation extraction in croatian, In LREC, pp. 657–662.
14. *Liu Tie-Yan* (2009), Learning to rank for information retrieval, Foundations and Trends in Information Retrieval, Vol. 3(3), pp. 225–331.
15. *Medelyan Olena, Frank Eibe, and Witten Ian H* (2009), Human-competitive tagging using automatic keyphrase extraction, In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Vol. 3, pp. 1318–1327.
16. *Pecina Pavel and Schlesinger Pavel* (2006), Combining association measures for collocation extraction, In Proceedings of the COLING/ACL on Main conference poster sessions, Association for Computational Linguistics, pp. 651–658.
17. *Ramisch Carlos* (2015), Evaluation of mwe acquisition. In Multiword Expressions Acquisition, Springer, pp. 105–125.
18. *Ramisch Carlos, Villavicencio Aline, and Boitet Christian* (2010), Mwetoolkit: a framework for multiword expression identification, In LREC.
19. *Vivaldi Jorge, Cabrera-Diego Luis Adrián, Sierra Gerardo, and Pozzi María* (2012), Using Wikipedia to validate the terminology found in a corpus of basic textbooks, In LREC, pp. 3820–3827.
20. *Wermter Joachim and Hahn Udo* (2006), You can't beat frequency (unless you use linguistic knowledge): a qualitative evaluation of association measures for collocation and term extraction, In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pp. 785–792.
21. *Wu Qiang, Burges Christopher J. C., Svore Krysta M., and Jianfeng Gao* (2010), Adapting boosting for information retrieval measures, Information Retrieval, Vol. 13(3), 254–270.

ВИДОВЫЕ ПАРЫ, СЕМАНТИЧЕСКАЯ ТЕОРИЯ И КРИТЕРИЙ МАСЛОВА

Урысон Е. В. (uryson@gmail.com)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Ключевые слова: русский глагол, видовая пара, критерий Маслова, лексическое значение, грамматическое значение, слабый семантический компонент, актуально-длительное значение несовершенного вида, полисемия

ASPECTUAL PAIRS, SEMANTIC THEORY AND MASLOV CRITERION

Uryson E. V. (uryson@gmail.com)

Russian Language Institute (Russian Academy of Sciences)

The paper deals with Russian aspectual pairs like umirat'—umeret', risovat'—narisovat' (but not obizhat'sa—obidet'sa). The imperfective verb in a pair designates a process or an action, while the perfective verb designates the "resulting event" completing the process / action. It is well known that in some diagnostic contexts the imperfective member of a pair substitutes it's perfective correlate and thus designates an event (Maslov criterion). It will be demonstrated that this substitution is due to certain semantic components in the lexical meaning of a verb. Fot this purpose the progressive meaning of imperfective verbs will be analysed. We will argue that the component 'the resulting event' is a part of meaning both of a perfective verb and the imperfective one in an aspectual pair. Status of this component in the lexical meaning of an imperfective verb will be discussed. Maslov's diagnostic contexts will be observed. A criterion for determing the imperfective correlate to a given perfective verb in some controversial cases (cf. est' / s'edat'—s'est') will be suggested in addition to Maslov criterion.

Key words: Russian verb, aspectual pairs, Maslov criterion, lexical meaning, grammatical meaning, weak semantic component, progressive meaning of an imperfective verb, polysemy

0. Объект и цель работы

Объект работы — русские видовые пары, в которых глагол НСВ обозначает действие или процесс, а глагол СВ выражает «достижение предела»; ср. *умирать — умереть, рисовать — нарисовать, ловить — поймать* и т. п. Таким образом, исключены из рассмотрения пары типа *обижаться — обидеться, возглавлять — возглавить* с иным семантическим соотношением глаголов СВ и НСВ.

Глагол рассматривается в каждом значении отдельно. Следовательно, видовая пара определяется для глагола в его конкретном значении, или, в соответствии с терминологией московской семантической школы, — для глагольной лексемы. Нас будет интересовать организация лексического значения глагола НСВ в сопоставлении с парным ему глаголом СВ.

Основная цель работы — показать, что эффективным инструментом верификации семантического описания глагола является критерий Маслова.

Поясним постановку задачи.

Семантическая теория стремится определить видовую пару исходя из значения глаголов НСВ и СВ: «чисто видовыми считаются пары глаголов, члены которых не выражают никаких семантических различий, кроме видовых» (Гловинская 1984: 24). В самом общем виде, глагол СВ (V1) и глагол НСВ (V2) образуют видовую пару, если лексическое значение глагола V1 тождественно лексическому значению глагола V2, а видовое значение СВ глагола V1 и видовое значение НСВ глагола V2 различаются строго определенным образом. Следовательно, для семантического определения видовой пары требуется располагать достаточно точными экспликациями лексической и видовой семантики глаголов. Однако такие экспликации предложены далеко не для всех глаголов, а имеющиеся семантические описания не всегда бесспорны.

Критерий Маслова тоже определяет видовую пару, но является операционным: он позволяет установить видовую пару, исходя из дистрибуции ее членов, т. е. парных глаголов СВ и НСВ, в определенных диагностических контекстах. По Ю. С. Маслову, глагол СВ (V1) и глагол НСВ (V2) образуют видовую пару, если при строго определенной трансформации контекста глагол СВ (V1) обязательно заменяется на глагол НСВ (V2). Например, глагол СВ автоматически заменяется на глагол НСВ «при переводе повествования из плоскости прошедшего времени в плоскость настоящего исторического» [Маслов 2004 / 1948: 76]. Ср. *Я съел суп и выкурил папиросу.* VS. *Я съедаю суп и выкуриваю папиросу.*

Обнаруженная Ю. С. Масловым дистрибуция членов видовой пары в диагностических контекстах имеет семантическую подоплеку: при описанной трансформации контекста «лексическая семантика глагола принципиально не должна подвергаться ни малейшему изменению» [Там же]. Именно поэтому «обратимость данного глагола СВ в тот или иной глагол НСВ при переводе повествования в плоскость исторического настоящего может служить надежным признаком парности этих двух глаголов, а необратимость — признаком того, что данные два глагола не составляют видовой пары» [Там же].

Мы попытаемся выявить лексическое значение парного глагола НСВ. На основе этого описания будет предложен критерий выбора глагола НСВ в качестве парного коррелята к глаголу СВ в некоторых спорных случаях, для которых критерий Маслова не дает однозначного решения.

Работа имеет следующую структуру.

В разделе 1 рассматривается актуально-длительное значение парного глагола НСВ. Демонстрируется, что один из основных компонентов этого значения требует обоснования, причем наличие данного компонента естественно верифицируется с помощью критерия Маслова. В разделе 2, на основе выводов раздела 1, обсуждается организация лексического значения парного глагола НСВ. В разделе 3 формулируется критерий выбора глагола НСВ в качестве члена видовой пары в некоторых спорных случаях. Предлагаемый критерий является дополнением критерия Маслова.

Ряд вопросов остается за пределами нашего исследования. Это, прежде всего, вопрос о статусе грамматической категории вида в русском языке (словоизменительная vs. словоклассифицирующая / словообразовательная) и вытекающий из этого вопрос о статусе в системе языка имперфективации vs. перфективации как способа образования видового коррелята. Обзор разных точек зрения по этой проблематике см. в работах [Горбова 2011; 2014; 2015]. В работе не затрагиваются и некоторые семантические проблемы, в частности, вопрос о «семантическом согласовании» значения приставки с остальной частью глагольной основы, а также вопрос о степени семантической близости приставочных производных друг к другу и к производящему глаголу [Janda 2007; Janda, Lyashevskaya 2011; Янда 2012; Janda et al. 2013].

1. Актуально-длительное значение НСВ и критерий Маслова

Хорошо известно, что актуально-длительное значение занимает привилегированное положение в иерархии частных значений НСВ. «Каждое частное значение НСВ требует для своей реализации того или иного контекста. Как пишет про частные видовые значения А. В. Бондарко, это “значения, выражаемые видом в сочетании с контекстом” (Грамматика 1980, т. 1: 604). Для актуально-длительного значения этот контекст минимальный, и в этом смысле оно является первичным. Для остальных значений толкование получается из толкования первичного значения как следствие изменения контекста» [Падучева 1996: 11]. Поэтому анализ семантики глагола НСВ естественно начать с актуально-длительного значения.

Рассмотрим толкование актуально-длительного значения НСВ, предложенное А. Вежбицкой [Wierzbicka 1967] для обозначений процессов, ср. *умирать*. По А. Вежбицкой, в толкование парного глагола НСВ, обозначающего процесс, в актуально-длительном значении входит указание на этот процесс, а кроме того импликация (упрощенно): если данный процесс дойдет до конца, то наступит новое состояние; это новое состояние обозначается парным глаголом СВ. Приведем толкование глагола *умирать* (в несколько модифицированном виде и с уточнением Т. В. Булыгиной [Булыгина 1983]):

(1) *Ян умирает* [актуально-длительное] ‘Ян близок к смерти и последовательно проходит состояния такого ряда состояний, что если он пройдет все состояния этого ряда, то наступит событие: Ян умер’.

В обобщенном виде (ср. [Падучева 1996: 18]):

(2) ‘Имеет место процесс P; если P не кончится преждевременно, то будет иметь место результирующее событие Q’.

Аналогичным образом можно толковать и глаголы НСВ, обозначающие действия, ср. *рисовать*. Действия предполагают цель, причем если действие не прекратится до определенного момента, то будет иметь место событие: цель достигнута. Это событие обозначается парным глаголом СВ. Ср. *Рисовал дерево, рисовал и постепенно его нарисовал*. Такие глаголы НСВ содержат следующий обобщенный компонент (ср. [Падучева 1996: 16–19]):

(3) ‘Субъект действует с целью P; если действие не прекратится преждевременно, то будет иметь место результирующее событие Q’.

Как отмечается, в частности, в книге [Гловинская 2001], действия типа *рисовать* предполагают «накопление результата» и этим сближаются с процессами типа *умирать, выздоравливать*. Другие парные глаголы НСВ предполагают «моментальное достижение цели», ср. *ловить* и его парный коррелят *поймать*: невозможно **Ловил бабочку, ловил и постепенно поймал ее*. Очевидно, однако, что компонент (3) присутствует и в толковании глаголов типа *поймать*.

Итак, в рассматриваемых видовых парах глагол СВ выражает результирующее событие — «достижение предела», а глагол НСВ содержит в толковании импликацию, указывающую, при каком условии предел будет достигнут, т. е. при каком условии будет иметь место результирующее событие. Покажем, что импликацию Вежбицкой выражают лишь те парные глаголы НСВ, которые удовлетворяют критерию Маслова.

Возьмем глагол *искать*. Семантически он сближается с глаголом *ловить*: оба глагола предполагают цель (*найти* и *поймать* соответственно), причем эта цель достигается моментально, без «накопления результата» (поэтому невозможно ни **постепенно <*медленно> найти монету*, ни **постепенно <*медленно> поймать зайца*). Естественно думать, что если импликация Вежбицкой справедлива для ситуации ловли и глагола *ловить*, то, наверно, она справедлива и для ситуации поиска и глагола *искать*. Действительно, ничто не мешает считать, что если субъект не прекратит свои действия, то рано или поздно он поймает того, кого ловит, или найдет тот объект, который ищет. Ср. известные строки: *Кто ищет, тот всегда найдет*. Если это так, то *искать* и *найти* — это такая же видовая пара, что и *ловить — поймать*. Такое решение принимается, например, в книге [Гловинская 2001: 107].

Обратимся, однако, к критерию Маслова. Глагол *ловить* в некоторых контекстах может обозначать событие ‘поймать’. Ср. *Он поймал фазана и получил*

вознаграждение [глагол СВ *поймать* обозначает событие] и *Он ловит фазана и получает вознаграждение* [настоящее историческое, глагол НСВ *ловить* обозначает событие ‘поймать’]. Глагол *ловить* может обозначать событие ‘поймать’ и в контексте сценической ремарки, ср. *Ловит птицу и уходит*. Наконец, этот глагол может обозначать событие в контексте многократности, ср. *Я много раз ловил таких птиц*. Что касается глагола *искать*, то он «ни при каких обстоятельствах не может иметь значения *найти*» [Маслов 1948: 313], почему и является непарным. Различное поведение глаголов *ловить* и *искать* в масловских контекстах требуется интерпретировать. Предлагаем следующую предварительную трактовку фактов.

Глагол *ловить* содержит в своей семантике указание на результирующее событие ‘поймать’. В частности, в актуально-длительном значении глагол *ловить* включает импликацию Вежбицкой, содержащую это указание. В настоящем историческом, в контексте сценической ремарки, в многократном контексте, а также в контексте, предполагающем общефактическое значение НСВ, это указание сохраняется — но уже в составе не импликации Вежбицкой, а какого-то другого выражения.

Что касается глагола *искать*, то в актуально-длительном значении он не выражает импликации (2). В другие частных значениях НСВ этот глагол тоже не указывает ни на какое событие. Поэтому естественно считать, что глагол *искать* не содержит в своей семантике указания на результирующее событие, и в этом заключается его кардинальное отличие от *ловить*.

Обратим, однако, внимание на то, что импликация (2), взятая безотносительно к конкретному глаголу, в точности соответствует нашему мироощущению, нашему общему представлению о действительности. Это представление «абстрагировано от конкретных ситуаций и связано с самыми абстрактными общими типами ситуаций или некоторыми общими принципами, отражающими мироощущение человека. Один из этих принципов утверждает: “Целенаправленное действие приводит к успеху”» [Санников 1989: 156]. Данный принцип был сформулирован В. З. Санниковым для описания употребления союзов *и* и *но*. Похожая «аксиома действительности» сформулирована в работе [Мартемьянов, Дорофеев 1983], посвященной анализу мира и человека на материале «Максим» Ларошфуко. Это совпадение неслучайно: если какая-то закономерность («аксиома») действительно является фрагментом общего для всех говорящих знания, то она и должна проявляться в самых разных текстах — от «Максим» Ларошфуко до высказываний с сочинительным союзом.

Замечательно, что представление о том, что целенаправленное действие приводит к успеху, закреплено в грамматике русского языка — в видовой системе глагола: многие видовые пары устроены так же, как пара *ловить — поймать*. Ср. *делать — сделать, рисовать — нарисовать, выгонять — выгнать, убивать — убить, прикреплять — прикрепить* и т. п. Но все же не все такие глаголы объединяются в пары: так, *искать* и *найти* видовой пары не образуют.

Рассмотрим некоторые аналогичные примеры.

Возьмем пары глаголов *убеждать — убедить* и *умолять — умолить*. Глаголы имеют сходное значение. Оба глагола НСВ (*убеждать* и *умолять*) обозначают речевое действие, цель которого — воздействовать на адресата так, чтобы

он сделал то, что хочет от него субъект. Ср. *Он убеждал ее уехать — Он умолял ее уехать*. Глаголы СВ *убедить* и *умолить* обозначают событие: «цель реализована». Ср. *Он убедил ее уехать — Он умолил ее уехать*. Казалось бы, перед нами две видовые пары, устроенные так же, как и пары *делать — сделать, вышивать — вышить* и т. п.

Однако по критерию Маслова видовую пару образуют лишь глаголы *убеждать — убедить*, но не *умолять — умолить*. Действительно, глагол НСВ *убеждать* в настоящем историческом может обозначать событие ‘убедить’. Ср.

(4) *Он убеждает ее уехать, а сам остается в Москве.*

(5) *Он приезжает в Москву и убеждает ее уехать с ним. (Они уезжают вместе).*

Глагол *убеждать* может обозначать событие и в контексте многократности, ср.

(6) *Он несколько раз убеждал ее ехать на красный свет. (В результате ей пришлось заплатить большой штраф)* [каждый раз имело место событие ‘он убедил ее ехать на красный свет’].

Что касается глагола *умолять*, то он в аналогичных (и других) контекстах обозначает только действие и не может обозначать результирующего события, т. е. достижения цели. Для того чтобы убедиться в этом, заменим в (5) глагол *убеждает* на *умоляет*, ср.

(7) *Он приезжает в Москву и умоляет ее уехать с ним.*

Высказывание (7), в отличие от примера (5), допускает только одно, «процессное» понимание: ‘Он приехал в Москву и умолял ее уехать с ним’. Данный пример не может значить ‘Он приехал в Москву и умолил ее уехать с ним’. Аналогичным образом не имеет «событийного прочтения» и следующее высказывание, в котором глагол *умолять* находится в контексте многократности:

(10) *Он несколько раз умолял ее ехать на красный свет.*

Следовательно, по критерию Маслова, глаголы *умолять — умолить* не образуют видовой пары, а семантически близкие им глаголы *убеждать — убедить* являются парными. По нашей интерпретации, глагол *убеждать* содержит в своем толковании импликацию Вежбицкой, а в глаголе *умолять* этого компонента значения нет. Разумеется, общий принцип «Целенаправленное действие приводит к успеху» справедлив и для ситуации ‘умолять’, однако соответствующая семантика в глаголе *умолять* не зафиксирована, при том что она является частью значения глагола *убеждать*.

Еще один пример — глаголы НСВ *течь* и *вытекать* в актуально-ддлительном значении в контекстах типа

(11) *Из рассохшейся бочки течет вода*.

(12) *Из рассохшейся бочки вытекает вода*.

Понятно, что примеры (11) и (12) ситуативно равнозначны. Очевидно также, что если описываемый процесс не прекратится, то вода из бочки вытечет. Иными словами, для обоих примеров справедлива импликация (2). Однако она является частью значения только глагола *вытекать*, но не глагола *течь*. Действительно, глагол *вытекать* в определенных контекстах может обозначать не процесс, а событие. Ср.

(13) *Вода из бочки вытекает, и нам приходится опять ее наполнять* [настоящее историческое].

(14) *Вода трижды вытекала из бочки* [контекст многократности].

Что касается глагола *течь*, то он в подобных контекстах может обозначать только процесс. Ср.

(15) *Вода из бочки течет, и нам приходится опять и опять ее наполнять* [настоящее историческое].

(16) *Вода трижды текла из бочки* [три раза имел место процесс].

Следовательно, глагол *течь* не содержит в своем значении указания на результирующее событие ‘вытечь’. Импликация (2) применительно к примеру (11) выражает лишь наше общее представление об устройстве мира.

2. Организация лексического значения выбранных глаголов

До сих пор мы говорили о том, входит или не входит указание на результирующее событие в значение глагола НСВ. Требуется уяснить, является ли этот компонент частью лексического значения глагола или же он относится к видовой семантике, т. е. входит в значение граммемы НСВ.

Обратим внимание на то, что указание на результирующее событие всегда входит в семантику глагола СВ: парный глагол СВ и его коррелят НСВ в данном отношении не различаются. Поэтому естественно считать, что указание на результирующее событие входит в лексическое значение обоих глаголов внутри пары.

Правда, в значении глагольной словоформы указание на событие всегда задействовано в видовой семантике. В глаголе НСВ в актуально-длительном значении оно участвует в импликации А. Вежбицкой (2), в глаголе НСВ в масловских контекстах это указание входит в состав какого-то другого выражения, в глаголе СВ оно участвует в видовой семантике СВ. Однако в лексическое

значение глагола это указание входит в «чистом виде», не осложненное видовым значением.

Поэтому мы не можем представить лексическое значение глагола каак таковое, оставаясь в рамках аналитического толкования глагольной словоформы: «истолковать глагольную основу безотносительно к виду невозможно» [Гловинская 2001: 56–57]. Тем не менее, с теоретической точки зрения вполне допустимо делать выводы о том, какие компоненты входят в лексическое значение глагола, а какие — в значение граммемы вида. Отметим, что указание на событие признается компонентом лексического значения НСВ Е. В. Падучевой [Падучева 1996: 86 сл.]; противоположной точки зрения придерживается М. Я. Гловинская [Гловинская 2001: 97].

Оказалось, что статус указания на результирующее событие в значении глагола НСВ может быть разным. Покажем это на примерах.

Возьмем глагол НСВ *читать*.

Глагол *читать* обозначает совершившееся событие, например, в настоящем историческом, ср.

(17) *Он прочитал статью и написал отзыв — Он читает статью и пишет отзыв.*

Результирующее событие, предполагаемое глаголом *читать*, таково: ‘субъект прочитал текст’.

Однако глагол *читать* нормально употребляется абсолютивно, ср.

(18а) *Весь день читает.*

(18б) *Что ты делаешь? — Читаю.*

В абсолютивном употреблении глагол *читать* сближается с масловскими глаголами «бесперспективного протекания» — он не указывает на «предел действия», т. е. не предполагает никакой результирующей ситуации. В частности, *читать* в абсолютивном употреблении не содержит импликации Вежбицкой. Следовательно, глагол *читать* указывает на результирующее событие лишь в определенных контекстах (например, при наличии прямого дополнения). В других контекстах это указание снимается.

Подобная организация лексического значения описана Ю. Д. Апресяном (для других классов слов) [Апресян 1990 / 1995: 486; 2004]. В соответствии с этим описанием, некоторые семантические компоненты внутри лексического значения могут подавляться контекстом — такие семантические компоненты называются слабыми, или неустойчивыми. С некоторой долей условности можно считать, что указание на результирующее событие в случае глагола *читать* снимается при абсолютивном употреблении. Тогда это указание является слабым компонентом лексического значения глагола *читать*.

Аналогичную семантическую организацию имеют и другие переходные глаголы, допускающие абсолютивное употребление. Ср. *шить, вышивать,*

писать, есть (суп), гладить (белье), стирать, штопать и т. п. Некоторые из них входят в так называемые «аспектуальные тройки»; ср. *читать — прочитывать — прочитать, есть — съедать — съесть* [Апреян 1995; Зализняк А., Микаэлян 2010]. Покажем, как можно использовать информацию о лексическом значении глагола для описания аспектуальных троек[1].

3. Дополнение к критерию Маслова

Рассмотрим поведение некоторых глаголов в масловских контекстах. Ограничимся тремя такими контекстами: настоящим историческим временем, настоящим в сценических ремарках и контекстом многократности[2]. Рассмотрим некоторые примеры и попытаемся их интерпретировать.

Возьмем глагол *есть (суп)*. В актуально-длительном значении он выражает импликацию Вежбицкой, т. е. указывает на возможное результирующее событие. Ср.

(19) *Что ты делаешь? — Ем суп* [если действие не прекратится преждевременно, будет иметь место событие: ‘субъект съел суп’].

В настоящем историческом глагол *есть* может обозначать уже не возможное, а совершившееся, реальное событие, ср.

(21) *Я съел суп, расплатился и пошел к выходу — Я ем суп, расплачиваюсь и иду к выходу.*

Следовательно, в контексте настоящего исторического импликация Вежбицкой в значении глагола *есть* заменяется другим выражением, в состав которого входит указание на совершившееся событие:

(22) ‘Имеет место результирующее событие Q’.

Однако это происходит не всегда: в следующем примере глагол *есть* в контексте настоящего исторического обозначает действие в его протекании и не указывает на реальное событие ‘субъект съел…’. Ср.

---

[1] Критерий Маслова может служить инструментом для выделения разных значений полисемичного глагола. В настоящей работе этот вопрос не рассматривается.

[2] Таким образом, мы вслед за А. В. Бондарко [Бондарко 1971], различаем два типа контекстов глагола — внешний и внутренний. Внешний контекст — это контекст глагольной словоформы, например, ее модификаторы, дополнения, однородные ей члены предложения и т. п. Внутренний контекст — это «окружение» собственно лексического значения глагола, т. е. граммемы, выражаемые глагольной словоформой, в частности, граммема времени или граммема НСВ в ее частном значении.

(23) *Я сидел в кафе, спокойно ел суп. (Вдруг раздался шум). — Я сижу в кафе, спокойно ем суп. (Вдруг раздается шум).*

Следовательно, импликация Вежбицкой в значении этого глагола в контексте настоящего исторического необязательно заменяется на выражение, содержащее компонент (22).

Не происходит этой замены и в контексте настоящего сценического. Ср.

(24) *Садится, ест суп* [действие в его протекании].

(25) *Садится, быстро ест суп, встает и уходит* [неясно, имело ли место событие 'субъект съел суп'].

В многократном контексте глагол *есть* тоже не указывает на реальное событие 'субъект съел суп', ср.

(26) *Я сегодня уже два раза ел суп.*

Итак, глагол *есть* в определенных контекстах содержит указание на результирующее событие: это событие либо потенциальное (импликация Вежбицкой; ср пример (19)), либо совершившееся (компонент (22), ср. пример (21)). Заметим, что глагол *есть* допускает абсолютивное употребление, ср. *Что ты делаешь? — Сижу, ем*. Указание на событие (потенцмальное) в таком контексте снимается.

Сравним теперь глаголы *есть* и *съедать*. Глагол *съедать* всегда указывает на событие 'съесть', причем оно имеет (имело или будет иметь) место безотносительно к каким-либо условиям. Ср.

(27) *Садится, съедает тарелку супа; Он уже два раза съедал весь суп.*

Указание на это событие входит и в значение глагола СВ *съесть*. Естественно считать, что данное указание является компонентом лексического значения обоих глаголов — *съедать* и *съесть*.

Сравним глаголы НСВ *есть* и *съедать* с глаголом СВ *съесть*. Лексическое значение глаголов *есть* и *съесть* содержит общий компонент — указание на событие. Но характеристика этого события, т. е. указание на его потенциальность (ср. импликацию Вежбицкой) или, наоборот, реальность (ср выражение (22)), является частью видового значения глагола. Что касается глаголов *съедать* и *съесть*, то у них общая часть значения больше: это указание на событие и на его реальность. По критерию Маслова оба глагола НСВ: и *есть*, и *съедать* — претендуют на роль видового коррелята глагола СВ *съесть*. Естественно считать, что в случае такой конкуренции в качестве члена видовой пары выбирается тот глагол НСВ, у которого больше общих компонентов значения с глаголом СВ. У глаголов *съедать* — *съесть* таких компонентов больше. К тому же в семантике глагола НСВ *есть* указание на событие является слабым

компонентом. Следовательно, видовую пару образуют глаголы *съедать — съесть*, но не *есть — съесть*.

Рассмотрим еще один пример. Вернемся к глаголу *читать*, исключив из рассмотрения абсолютивные контексты.

В актуально-длительном значении глагол *читать* выражает имплика-цию Вежбицкой, т. е. указывает на потенциальное событие. Ср.

(28) *Что ты делаешь? — Читаю статью* [если действие не прекратится преждевременно, будет иметь место событие: ‘субъект прочитал статью’].

В настоящем историческом глагол *читать* может обозначать совершив-шееся, т. е. реальное событие ‘субъект прочитал статью’. Ср.

(29) *Он прочитал статью и написал отзыв — Он читает статью и пишет отзыв.*

Однако это имеет место не всегда: в следующем примере глагол *читать* в контексте настоящего исторического обозначает действие в его протекании, т. е. указывает на потенциальное, а не на реальное событие. Ср.

(30) *Я сидел, спокойно читал статью. (Вдруг зазвонил телефон). — Сижу спокойно читаю статью. (Вдруг звонит телефон).*

Аналогичным образом ведет себя глагол *читать* в контексте настоя-щего сценического. Ср.

(31) *Садится, быстро читает письмо, встает и уходит* [имеет место событие ‘субъект прочитал письмо’].

(32) *Садится, читает письмо* [действие в его протекании; указание на потенциальное событие].

Что касается многократного контекста, то в нем глагол *читать* нор-мально указывает на совершившееся, реальное событие, ср.

(33) *Я несколько раз читал эту статью, но так ничего и не понял; Я много раз читал этот роман.*

Итак, глагол НСВ *читать* не в абсолютивном употреблении со-держит указание на результирующее событие: либо потенциальное, либо совершившееся.

Сравним *читать* и *прочитывать*. Последний глагол всегда указывает на реальное событие. Ср.

(34) *Он быстро <спокойно> прочитывает статью.*

Указание на реальное событие входит в значение глагола СВ *прочитать*: это общий компонент лексического значения глаголов *прочитывать* и *прочитать*.

Глаголы *читать* и *прочитать* тоже имеют общий компонент — указание на событие. Однако модальная характеристика этого события, т. е. указание на его потенциальность или, наоборот, реальность, является частью не лексического, а видового значения глагола.

По критерию Маслова оба глагола НСВ: и *читать*, и *прочитывать* — претендуют на роль видового коррелята глагола СВ *прочитать*. Однако у глаголов *прочитывать* и *прочитать* общая часть значения больше: это указание на событие и на его реальность. Поэтому видовую пару образуют глаголы *прочитывать — прочитать*, но не *читать — прочитать*.

Отметим, что глаголы НСВ *съедать* и *прочитывать* имеют дефектный набор частных видовых значений — в нем отсутствует актуально-длительное значение. На наш взгляд, дефектность набора форм не имеет отношения к определению грамматической категории.

Заметим, что глагол *читать* указывает на совершившееся событие в большем числе контекстов, чем глагол *есть*. Поэтому *читать* в каком-то смысле ближе к *прочитать*, нежели *есть* к *съесть*. Исследование различной степени близости глаголов СВ и НСВ внутри видовой (или предположительно видовой) пары представляет собой отдельную задачу.

Итак, если на роль видового коррелята к глаголу СВ претендуют два глагола НСВ, то в качестве члена видовой пары выбирается тот глагол НСВ, у которого больше семантическая общность с глаголом СВ. В рассмотренных примерах такой глагол НСВ, по-видимому, указывает на совершившееся событие во всех масловских контекстах.

Однако теоретически можно предположить, что из двух претендентов НСВ на роль видового коррелята глагола СВ ни один не обозначает совершившегося события во всех диагностических контекстах. Тогда членом видовой пары естественно считать тот глагол НСВ, который указывает на совершившееся событие в большем числе контекстов. Предлагаемое правило выбора глагола НСВ на роль видового коррелята к глаголу СВ может служить дополнением к критерию Маслова.

Заключение

Продемонстрировано, что операционный критерий Маслова позволяет определить, входит ли в лексическое значение глагола НСВ указание на событие, а также установить, является ли этот компонент значения слабым или сильным. В спорных случаях, когда на роль видового коррелята глагола СВ претендует более одного глагола, естественно выбирать тот глагол НСВ, в лексическом значении которого больше общих элементов с глаголом СВ. Дальнейшая задача состоит в том, чтобы рассмотреть с этой точки зрения разные аспектуальные тройки. Предлагаемый подход может дополнить детальное описание видовых троек, предлагаемое в работе [Зализняк Анна, Микаелян 2010].

Тот факт, что указание на потенциальное событие (импликация Вежбицкой) входит не во все глаголы НСВ, обозначающие целенаправленное действие, свидетельствует о том, что семантическое противопоставление ‘действовать с целью’ — ‘цель реализована’ не вполне закреплено в системе вида. Это может быть следствием относительной молодости грамматической системы вида (которая проявляется и в морфологии: в богатстве и сложных правила выбора средств, оформляющих видовое противопоставление).

Работа выполнена при финансовой поддержке РНФ (проект 16-18-02054).

Литература

1. *Apresyan Yu. D.* (2007), Lexical semantics [Leksicheskaya semantika], Nauka, Moscow.
2. *Apresyan Yu. D.* (1995), Interpretation of excess aspectual paradigms in an explanatory dictionary [Traktovka izbytochnykh aspektulnykh paradigm v tolkovom slovare], Apresyan Yu. D., Selected works, V. II, Yazyki russkoy kul/yury, Moscow, pp. 102–113.
3. *Bondarko A. V.* (1971), Aspect and tense of Russian verb, Prosveshcheniye, Moscow.
4. *Bulygina T. V.* (1983), Classes of predicates and aspect characteristics of an utterance [Klassy predikatov I aspektualnaya kharakteristika vyskazyvaniya], Aspect and tense meanings in slavic languages: Workshop proceedings of the Comission on studying grammatical structure of Slavic languages, Moscow, pp. 20–39.
5. *Glovinskaya M. Ya.* (1984), On the notion of a pure ashectual pair [K poniatiyu chisto vidovoy pary], Problems of structural linguistics 1982, Nauka, Moscow, pp. 24–34.
6. *Glovinskaya M. Ya.* (1984), Polysemy and synonymy in aspect-tense system of Russian verb [Polisemiya i sinonimiya v vido-vremennoy sisteme russkogo glagola], Azbukovnik — Russkie slovari, Moscow.
7. *Gorbova E. V.* (2011), Aspectual twoness of Russian verb: problems and solutions [Vidovaya parnost’ russkogo glagola: problemy I resheniya], Questions of linguistics, № 4, pp. 20–45.
8. *Gorbova E. V.* (2014), Once more about aspect speciation of Russian verb: on inflexion interpretation of aspect [Eshcho raz o vidoobrazovanii russkogo glagola: k slovoizmenitel’noy traktovke vida], Russian Linguistics, 2014. Vol. 38. № 2.
9. *Gorbova E. V.* (2015), Aspect speciation of Russian verb: prefixation or suffixation? [Vidoobrazovaniye russkoogo glagola: prefiksatsiya ili suffiksatsia?], Questions of linguistics, № 1. pp. 7–38.
10. *Janda L.* (2007), Aspectual clusters of Russian verbs, Studies in Language, Vol. 31, No. 3, pp. 607–648.
11. *Janda L, Lyashevskaya O.* (2011), Aspectual pairs in the Russian national corpus, Scando-Slavica, Vol. 57, No 2, pp. 201–215.
12. *Janda L.* (2012), Russian prefixes as a system of verb classifiers [Russkie pristavki kak sistema glagolnykh klassifikatorov], Questions of linguistics, № 6.

13. *Janda L., Endersen A., Kuznetsova J., Lyashevskaya O., Makarova A., Nesset T., Sokolova S.* (2013), Why Russian aspectual prefixes aren't empty. Prefixes as verb classifiers, Slavica Publishers Indiana U., Bloomington.
14. *Koschmieder E.* (1962), An outline for study of Polish verb aspects [Ocherk nauki o vidakh polskogo glagola], Questiona of verb aspect, Nauka, Moscow, pp. 159–161 [first ed.: Koschmieder E. Nauka o aspektach czasownika polskiego w zarysie, Wilno, 1934. S. 97–102].
15. *Martemyanov Yu. S., Dorofeev G. V.,* (1983), A trial for terminologisation of literary language vocabulary: about world of vanity according to Fr. de La Rochefoucauld [Opyt terminologisatsii obshcheliteraturnoy leksiki: o mire tshcheslaviya po F. De Laroshfuko], Questions of cybernetics: logic of reasoning and it's modelling, Moscow, pp. 38–103.
16. *Maslov Yu. S.* (2004), Verb aspect and lexical meaning in modern Russian literary language [Vid i leksicheskoye znacheniye glagola v sovremennom russkom literaturnom yazyke], Maslov Yu. S., Selected works: Aspectology. General linguistics, Yazyki slavyanskikh kul'tur, Moscow, pp. 71–90.
17. *Maslov Yu. S.* (2004), System of concrete aspectual meanings and types of opposition of perfective and imperfective aspects [Sistema chastnykh vidovykh znacheniy I tipy protivopostavleniy sovershennogo i nesovershennogo vidov], Maslov Yu. S., Selected works: Aspectology. General linguistics, Yazyki slavyanskikh kul'tur, Moscow, pp. 96–110.
18. *Zalizniak Anna, Mikaelian I.* (2010), On status of aspectual triples in the Russian aspectual system, [O meste aspektual'nykh troek v vidovoy sisteme russkogo yazyka], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2010" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2010"], Iss. 9 (16), Bekasovo, pp. 130–136.
19. *Paducheva E. V.* (1996), Semantic researches (Semantics of aspect and tense in Russian; Semantics of narrative) [Semanticheskiye issledovaniya (Semantika vida i vremeni v russkom yazyke; Semantika narrativa)], Eazyki russkoy kul'yury, 1996.
20. *Sannikov V. Z.* (1989), Russian coordinating constructions, Nauka, Moscow.
21. *Wierzbicka A.* (1967), On the Semantics of the Verbal Aspect in Polish. in To Honor Roman Jakobson. Essays on the Occasion of his Seventieth Birthday, The Hague — Paris: Mouton, pp. 2231–2243.

СРАВНЕНИЕ КОРПУСНОГО И ЭКСПЕРИМЕНТАЛЬНОГО МЕТОДА НА ПРИМЕРЕ ИССЛЕДОВАНИЯ СИНТАКСИЧЕСКИХ СВОЙСТВ ЭНКЛИТИКИ *ЖЕ*

Валова Е. А. (dunya_v@yahoo.com)
НИУ ВШЭ, Москва, Россия

Слюсарь Н. А. (slioussar@gmail.com)
НИУ ВШЭ, Москва; СПбГУ, Санкт-Петербург, Россия

Ключевые слова: русский язык, фразовые энклитики, закон Ваккернагеля, корпусный метод, экспериментальный метод

COMPARING CORPUS AND EXPERIMENTAL APPROACHES: A STUDY OF SYNTACTIC PROPERTIES OF THE RUSSIAN ENCLITIC *ŽE*

Valova E. A. (dunya_v@yahoo.com)
HSE, Moscow, Russia

Slioussar N. A. (slioussar@gmail.com)
HSE, Moscow;
St. Petersburg State University, St. Petersburg, Russia

Corpus and experimental approaches in linguistics are often seen as incompatible, and there are very few studies of grammatical phenomena that rely on both of them, without one or the other being subsidiary. In this paper, we would like to show that they are complimentary and can be fruitfully combined on the example of Russian phrasal enclitic *že*. We analyze various factors influencing its position in the sentence, in particular, whether it obeys Wackernagel's law, which applied to phrasal enclitics in Old Russian.

Data from the National Russian corpus show that *že* appears in the strict Wackernagel's position in the absolute majority of cases in the main subcorpus and the newspaper subcorpus, while the subcorpus of spoken Russian exhibits more variation. Corpus data allow tracing diachronic tendencies and identifying several factors (primarily, the semantics of *že*). Experimental data let us estimate the role of these and other factors on a carefully balanced set of examples. Apart from syntactic and semantic factors, the age and educational level of participants was demonstrated to influence the results.

Keywords: Russian, phrasal enclitics, Wackernagel's law, corpus approach, experimental approach

1. Введение

В середине XX в. в лингвистике началось несколько важных дискуссий, которые в том или ином виде продолжаются до сих пор. Одна из них касалась того, какие источники данных должны использоваться в исследовании. Ее инициировал Н. Хомский, который обрушился с критикой на корпусный подход и предложил использовать вместо него интроспективные суждения о грамматической правильности тех или иных примеров (grammaticality judgments) ([Chomsky 1957, 1962, 1965] и далее). Эта идея сыграла ключевую роль в развитии экспериментального подхода: многие авторы, опасаясь полагаться исключительно на свою интуицию, стали опрашивать большие группы носителей языка.[1]

С тех пор и корпусные, и экспериментальные методы в лингвистике претерпели множество изменений и значительно усовершенствовались. Тем не менее, антагонизм между их последователями во многом сохранился. Продолжают появляться работы, предлагающие новые аргументы против одного из этих подходов, а также отвечающие на высказанную ранее критику (см. например, [McEnery, Wilson 2001; Riemer 2009; Sampson 1992; Schütze 1996; Tremblay 2005]).

Мы полагаем, что, во всяком случае, на данном этапе развития лингвистики эта дискуссия неплодотворна, потому что никто не заставляет исследователя ограничиваться одним подходом: их можно рассматривать не как взаимоисключающие, а как взаимодополняющие. Однако для того, чтобы убедиться в этом, нужны конкретные примеры. Между тем, подавляющее большинство исследований либо полагается только на один из этих источников данных, либо, используя оба, рассматривает один из них сугубо как вспомогательный (см., например, обзор в [Guilquin, Gries 2009]). В данной работе мы хотим внести свою лепту в решение этой проблемы, продемонстрировав, как корпусные и экспериментальные данные могут дополнить друг друга, на примере исследования синтаксических свойств фразовой энклитики *же*.

Клитиками называют слова, не имеющие собственного ударения и подчиняющиеся акцентуации предшествующего или последующего слова (например, энклитика *бы* в *сказал бы,* проклитика *на* в *на полке*). Некоторые энклитики являются фразовыми, то есть относятся не к конкретному слову, а к предложению в целом (например, *Я же говорил об этом! Говорил ли я об этом?*). В древнерусском языке, как и в праиндоевропейском, расположение таких энклитик подчинялось закону Ваккернагеля [Wackernagel 1892]: они примыкали к первому фонетическому (полноударному) слову фразы. Однако затем его действие ослабло (ср. [Зализняк 2008]).

Мы провели корпусное и экспериментальное исследование, чтобы оценить, насколько оно ослабло, на примере энклитики *же*, которая лучше всего сохранила свои ваккернагелевские свойства [Зализняк 2008]. Подробному описанию различных результатов этих исследований будет посвящена другая

---

[1] Один из рецензентов отметил, что терминологически правильнее говорить о противопоставлении корпусного метода и метода опроса, но мы будем придерживаться устоявшейся терминологии.

работа[2]. Здесь же мы хотели бы, осветив вкратце основные моменты, сфокусироваться на вопросах, связанных с совместным использованием корпусных и экспериментальных данных.

Ослабление закона Ваккернагеля во многих языках выражается в том, что он начинает действовать в так называемой нестрогой формулировке, согласно которой фразовые энклитики могут располагаться как после первого полноударного слова, так и после первой синтаксической составляющей (ср., например, [Spencer, Luís 2012]). Подобную вариативность можно наблюдать и для фразового *же*, и в данной работе она была исследована на материале именных групп. Возможные варианты размещения энклитики мы будем называть соответственно первой позицией, или строгой позицией Ваккернагеля, как в примере (1), и второй, или нестрогой ваккернагелевской позицией, как в примере (2).

(1) *Но глупый плачет от бедства, а умный ищет средства.* ***Татьяна же Акимовна*** *была женщина умная.* [Н.Г. Помяловский. Махилов (1855)]

(2) *ФИФА унифицировала свой регламент по переходам еще в прошлом году.* ***У национальных федераций же*** *есть время на его адаптацию.* [Мутко: В российском футболе появится статус свободного агента // Советский спорт, 2006.10.11]

2. Корпусное исследование

Корпусное исследование было проведено на материале Национального корпуса русского языка (НКРЯ, www.ruscorpora.ru). Так как примеров с фразовым *же* в НКРЯ очень много, мы остановились на предложениях, где носителями энклитики являются два типа именных групп: комплекс имен собственных, как в примере (1) выше, и сочетание прилагательного и существительного, как в примере (2). Набор факторов, которые могут действовать в первой группе примеров, ограничен (в частности, не идет речи о большей или меньшей семантической связности двух слов внутри именной группы), их число невелико. Мы ожидали, что в них особенно ярко проявится тенденция к ослаблению закона Ваккернагеля. Вторая группа примеров намного более разнородна и обширна.

После отсева многочисленных примеров с локальным *же*[3], в нашу итоговую выборку вошло около пятнадцати тысяч предложений. Для примеров из основного подкорпуса НКРЯ мы проследили, как менялось число предложений с *же* во второй позиции с течением времени. С середины 19 в. наблюдается рост их доли в обеих рассматриваемых группах (до этого подобные примеры единичны), применение логистической регрессии показывает, что эта тенденция

---

[2] Кроме того, предварительные результаты корпусного исследования были представлены в [Валова 2014a, b].

[3] В этих примерах энклитика относится не к предложению в целом, а к предшествующему слову (например, *такой же ответ*, *на следующий же день*).

статистически значима ($p < 0{,}01$). Однако, как показывают обобщенные данные в Таблице 1, если в случае с именами собственными картина изменяется заметно, то в случае с прилагательными и существительными число таких примеров остается настолько незначительным, что они совершенно незаметны на общем фоне.

Таблица 1. Частотность примеров с *же* в первой и второй позициях: основной корпус

Временной интервал	Имена собств.		Прил. и сущ.		Общее кол-во словоупотреблений
	поз. 1	поз. 2	поз. 1	поз. 2	
1851–1900	155	4 (2,5%)	1690	5 (0,3%)	41 107 495
1901–1950	131	17 (11,5%)	1964	17 (0,9%)	60 237 302
1951–2000	103	16 (13,5%)	1632	22 (1,3%)	73 970 923
2001–н.вр.	37	19 (33,9%)	1551	35 (2,2%)	283 906 971

Более внушительную долю примеров с *же* во второй позиции можно найти, сравнивая данные основного, газетного и устного подкорпусов НКРЯ. Если взять все вошедшие в нашу выборку предложения с прилагательными и существительными, мы увидим, что в основном подкорпусе таких примеров 85 (1,1%), в газетном — 166 (3,2%), а в устном — 25 (25,5%) (в устном и газетном подкорпусе все примеры относятся к самому позднему периоду, в основном самые ранние — к началу 18 в.). Все три подкорпуса значимо отличаются друг от друга ($p < 0{,}01$ согласно критерию хи-квадрат с поправкой Бонферрони).

Кроме того, корпусное исследование позволило выявить ряд факторов, которые влияют на позицию энклитики. Прежде всего, это ее значение. Рассмотрению различных значений *же* посвящена обширная литература (например, [Киселев 1976; Paducheva 1987; Бонно, Кодзасов 1998]). Опираясь на нее, в нашей работе мы выделяли следующие значения: противительное, усилительное, пояснительное и присоединительное.

Во всех группах примеров во всех подкорпусах более высокий процент предложений с нестрогой ваккернагелевской позицией наблюдается для пояснительного и усилительного *же*, самый низкий — для противительного. Однако при этом примеров с противительным *же* на порядок больше, чем всех прочих. Примеров с присоединительным *же* слишком мало, чтобы делать какие-либо выводы. Также мы отметили, что в предложениях с отрицанием *же* в абсолютном большинстве случаев занимает строгую ваккернагелевскую позицию, а в предложениях, где прилагательное и существительное образуют устойчивое словосочетание (коллокацию), *же* чаще оказывается во второй позиции. Были описаны и некоторые другие тенденции. В целом же мы пришли к выводу, что, хотя случаи нарушения строгого закона Ваккернагеля в рассмотренных материалах не единичны, ослабление его действия далеко не так значительно, как можно бы было ожидать.

3. Экспериментальное исследование

Мы провели два эксперимента с использованием разных методик. В первом респонденты должны были разместить *же* в предложениях, используя одну из двух предложенных позиций. Во втором респондентам было предложено оценить предложения по шкале от 1 до 5, где *1* — «предложение звучит плохо, я бы так никогда не сказал», а *5* — «предложение звучит хорошо, я сам мог бы так сказать»[4].

Основной минус первого метода заключается в том, что, если испытуемые считают приемлемыми (или, наоборот, неприемлемыми) оба варианта, мы не получим об этом информации — только информацию о том, какой из двух вариантов кажется (несколько) лучше. Зато он дает сведения о предпочтениях в максимально простом и удобном для анализа виде. Основным минусом второго метода можно считать то, что нельзя предъявить одному испытуемому оба варианта предложения (с *же* в первой и во второй позициях), так как повторять стимулы в эксперименте не принято. Поэтому во втором эксперименте было два экспериментальных листа (половина испытуемых видела то или иное предложение с *же* в первой позиции, половина — с *же* во второй). Анкетирование проводилось с использованием форм Google (www.google.com/intl/ru_ru/forms/about/).

В первом эксперименте приняло участие 214 человек в возрасте от 14 до 80 лет, из них 53 мужчины и 161 женщина. Второй прошли 181 человек в возрасте от 16 до 80 лет, из них 42 мужчины и 139 женщин. Уровень образования респондентов варьировал от незаконченного среднего до наличия ученой степени.

Чтобы можно было сопоставить результаты корпусного и экспериментального исследования, носителем энклитики в стимульных предложениях была сделана именная группа с прилагательным и существительным. Ниже перечислены различные группы стимулов (их сокращенные названия, описание и примеры)[5]. Возможные позиции энклитики, которые рассматривались в данной работе, обозначены скобками, для термина *именная группа* используется сокращение *ИГ*.

- Прот. *же* (не ИГ): *же* в противительном значении, противопоставлены клаузы.

  Я три года деньги коплю, деревянный (_) *домик* (_) *всё дорожает и дорожает.*

- Прот. *же* (ИГ): *же* в противительном значении, противопоставлены ИГ.

---

[4] В настоящее время в экспериментах обычно используются не бинарные оценки («правильно/неправильно»), а шкалы, которые позволяют отслеживать более тонкие различия.

[5] Кроме того, была создана группа примеров, в которых именной группой — носителем клитики является устойчивое выражение, или коллокация, а также аналогичные сочетания, демонстрирующие меньшую связанность, однако мы не будем рассматривать их в данной работе.

Ты как в воду глядел. Кирпичные многоэтажки никого не впечатлили, деревянный (_) *домик* (_) *вызвал всеобщий энтузиазм.*

- Прот. *же* (прил.): *же* в противительном значении, противопоставлены прилагательные внутри ИГ.

  Ты как в воду глядел. Кирпичные многоэтажки никого не впечатлили, деревянный (_) *домик* (_) *вызвал всеобщий энтузиазм.*

- Поясн. *же: же* в пояснительном значении.

  Сфотографируй это! Деревянный (_) *домик* (_) *такой красивый.*

- Поясн. *же* (прот. ИГ): *же* в пояснительном значении, при этом в предложении есть противопоставление ИГ.

  Всё вышло, как ты хотел. Деревянный (_) *домик* (_) *останется, а бетонные громадины снесут.*

- Поясн. *же* (отр.): *же* в пояснительном значении, в предложении есть отрицание.

  Чему ты радуешься? Не деревянный (_) *домик* (_) *тут будет!*

- Усил. *же* (констр.): *же* в усилительном значении, конструкция с вынесением ремы.

  Красивый (_) *домик* (_) *купили наши соседи!*

- Присоед. *же: же* в присоединительном значении.

  Чтобы что-то построить, надо было снести один деревянный домик или купить еще земли. Деревянный (_) *домик* (_) *охранялся как памятник, и снести его было сложно.*

Таким образом, у нас было восемь групп стимулов, в каждой из которых было по три примера (то есть всего 24 предложения). В этих предложениях *же* появлялось в разных значениях. Кроме того, внутри группы с противительным *же* рассматривались примеры с разным фокусом контраста (в которых противопоставлены именные группы — носители клитики, входящие в них прилагательные или клаузы в целом), а в группе примеров с пояснительным *же* были примеры с отрицанием и с противопоставлением именных групп (мы хотели оценить, какую роль играет само наличие в предложении противопоставления, а какую — значение *же*). Пример с усилительным *же* содержал особую конструкцию с вынесением ремы в начало предложения. Также в эксперимент

были включены 64 филлера: предложения с частицами *бы* и *не*, предполагающие выполнение тех же заданий, что и стимулы, но на другом материале (например, в предложение «Я всю жизнь ждала человека, который (_) понял (_) меня» надо было вставить частицу *бы* или оценить его с частицей в разных позициях).

Результаты экспериментов обобщены в Таблице 2. Первая колонка отражает процент выбора нестрогой ваккернагелевской позиции в первом эксперименте. Средняя оценка первой и второй позиций, а также их разница приведены на основании второго эксперимента.

Таблица 2. Результаты двух экспериментов

Тип стимула	**% выбора поз. 2**	**Оценка поз. 1**	**Оценка поз. 2**	**Разница**
Прот. *же* (не ИГ)	58,9%	3,6	4,0	−0,4
Прот. *же* (ИГ)	33,2%	4,4	3,9	0,5
Прот. *же* (прил.)	6,3%	4,7	3,3	1,4
Поясн. *же*	85,3%	2,4	3,9	−1,5
Поясн. *же* (прот. ИГ)	51,9%	2,8	3,1	−0,4
Поясн. *же* (отр.)	14,3%	4,4	3,5	0,9
Усил. *же* (констр.)	3,7%	4,5	2,6	1,9
Присоед. *же*	57,2%	3,9	4,2	−0,3

Сперва рассмотрим результаты первого эксперимента. Используя метод логистической регрессии, мы оценили вероятность ответа «1» или «2» в зависимости от различных характеристик предложений. Три типа предложений с противительным *же* значимо отличаются друг от друга ($p < 0{,}01$), что объясняется различным фокусом контраста. Значимо отличаются друг от друга и три типа предложений с пояснительным *же* ($p < 0{,}01$): процент выбора второй позиции снижается при появлении противопоставления и особенно при наличии в предложении отрицания. Предложения с противительным и пояснительным *же*, включающие противопоставление именных групп, значимо отличаются друг от друга ($p < 0{,}01$).

В целом мы полагаем, что роль значения энклитики правильнее оценивать на материале корпусных данных, поэтому в эксперименте стремились не сравнить друг с другом все возможные значения, а скорее удостовериться, что этот фактор играет роль, и посмотреть, как он будет взаимодействовать с другими. Однако в корпусе было слишком мало предложений с присоединительным *же*, поэтому мы сравнили примеры с пояснительным, присоединительным и противительным *же* на материале экспериментальных данных (из рассмотрения были исключены предложения с отрицанием и противопоставлением прилагательных, так как в них действуют мощные независимые факторы). В результате было выявлено то же распределение, которое наблюдается в корпусе, и различия между группами оказались статистически значимыми ($p < 0{,}01$).

На примере усилительного *же* мы планировали оценить не фактор значения (таких примеров достаточно в корпусе), а роль особой конструкции, для

которой характерен определенный интонационный контур и определенная интерпретация (именная группа является ремой, вынесенной в начало предложения). Результаты показали, что в таких предложениях процент выбора нестрогой ваккернагелевской позиции был самым низким (в частности, эта группа значимо отличается от группы, содержащей отрицание, $p < 0{,}01$).

О результатах второго эксперимента мы скажем очень кратко. Прежде всего, заметим, что результаты двух экспериментов согласуются друг с другом: если какой-то вариант чаще выбирают в первой части, у него всегда выше средний балл во второй. Учитывая это, мы не будем повторять для второй части приведенные выше сравнения, а остановимся на других моментах. Прежде всего, бросается в глаза, что, даже при наличии явных предпочтений, очень часто оба варианта предложения получают достаточно высокие оценки. Это верно для всех примеров с противительным *же* (несколько в меньшей степени при противопоставлении прилагательных), для присоединительного *же* и, что было для нас неожиданным, даже для предложений с пояснительным *же*, в которых есть отрицание. Как нам кажется, такую картину можно считать дополнительным свидетельством ослабления закона Ваккернагеля в его строгой формулировке.

Относительно низкие оценки получили предложения с пояснительным *же* в первой позиции и с усилительным *же* в рамках особой конструкции во второй — судя по всему, эти два варианта действительно режут респондентам слух. Кроме того, невысокие оценки получили оба варианта предложений с пояснительным *же*, включавшие противопоставление именных групп. Видимо, они в целом не понравились испытуемым (например, потому, что в них лучше смотрелось бы *ведь*, а не *же*), поэтому данные по этим предложениям надо рассматривать с осторожностью.

Кроме того, на материале первого эксперимента мы изучили, влияют ли на ответы различные социолингвистические характеристики респондентов. Мы выяснили, что уровень образования и особенно возраст являются значимыми факторами ($p = 0{,}04$ и $p < 0{,}01$ соответственно), а пол не является ($p = 0{,}80$). Более молодые и/или менее образованные респонденты чаще выбирали для энклитики вторую позицию.

4. Заключение

В данной работе мы представили результаты корпусного и экспериментального исследования синтаксических свойств энклитики *же*. Прежде всего, отметим, что многие тенденции проявились и в корпусных, и в экспериментальных данных. Например, в обоих исследованиях реже всего оказывается в нестрогой ваккернагелевской позиции противительное *же*, а чаще всего — пояснительное, в обоих проявилась роль наличия в предложении отрицания. Это можно считать лишним доказательством того, что эти два метода не противоречат друг другу.

В определенных случаях эти методы дополняют друг друга, что представляется особенно ценным. Так, корпусное исследование лучше подходит для того, чтобы судить о роли значения энклитики для ее расположения

в предложении — в этом случае важно опираться на большую выборку разнообразных примеров, порожденных в естественных условиях. А используя экспериментальный метод, легче на однородном материале оценить влияние различных формальных факторов: наличия отрицания, противопоставления, особой конструкции, изменение фокуса контраста. Корпус дает представление о вариации в текстах разных эпох и жанров, эксперимент — о вариации, связанной с социолингвистическими факторами (возраст, пол. уровень образования респондентов).

Наконец, нельзя не заметить и существенные различия между результатами корпусного и экспериментального исследования. В основном подкорпусе Национального корпуса русского языка примеры с прилагательными и существительными, где *же* находится во второй позиции, практически незаметны на общем фоне, в то время как в эксперименте таких ответов чуть меньше половины. Прежде всего, отметим, что расхождения оказываются далеко не такими разительными, если сравнить результаты эксперимента с данными устного подкорпуса. Таким образом, отчасти мы имеем дело не с различиями между методами, а с различиями между жанрами.

Определенные расхождения, безусловно, связаны и с различиями между методами. Однако, как нам кажется, неверно будет говорить, что один из методов подводит нас ближе к истинной картине, чем другой: скорее они позволяют посмотреть на нее с разных сторон. Например, мы показали в эксперименте, что есть конструкции, прежде всего, примеры с пояснительным *же*, в которых подавляющее большинство опрошенных носителей языка уверенно выбирает для энклитики нестрогую ваккернагелевскую позицию. Однако нам нужен корпус для того, чтобы понять, что *же* в пояснительном значении используется на порядок реже, чем в противительном. Соответственно, роль таких конструкций может быть велика (всякая глобальная перестройка системы начинается с ограниченного класса примеров и затем захватывает всё новые и новые), но, во всяком случае, на данном этапе развития русского языка, они мало заметны в текстах в силу общей низкой частотности. В связи с этим мы полагаем, что наиболее полное представление о том или ином языковом явлении можно получить только путем использования обоих методов.

Часть представленных в данной статье исследований была выполнена при поддержке гранта РНФ № 16-18-02 071.

Литература

1. *Бонно К., Кодзасов С. В.* (1998), Семантическое варьирование дискурсивных слов и его влияние на линеаризацию и интонирование (на примере частиц «же» и «ведь»), Дискурсивные слова русского языка: опыт контекстно-семантического описания, Киселева К., Пайар Д. (ред.), Метатекст, Москва, с. 382–446.
2. *Валова Е. А.* (2014), Синтаксические свойства энклитической частицы *же* в диахроническом аспекте: корпусное исследование. // Научно-техническая информация. Серия 2. Информационные процессы и системы, № 10, с 31–36.

3. *Валова Е. А.* (2014). Синтаксические свойства русской энклитической частицы *же*. // Вестник Православного Свято-Тихоновского гуманитарного университета. Серия 3. Филология. №III:4 (39), с. 16–33.
4. *Киселев И. А.* (1976), Частицы в современных восточнославянских языках, БГУ, Минск.
5. *Зализняк А. А.* (2008), Древнерусские энклитики, Языки славянских культур, Москва.
6. *Chomsky N.* (1957), Syntactic Structures, Mouton and Co., The Hague.
7. *Chomsky N.* (1962), Explanatory Models in Linguistics, Logic, Methodology, and Philosophy of Science, Stanford University Press, Stanford, Calif., pp. 528–550.
8. *Chomsky N.* (1965), Aspects of the Theory of Syntax, The M. I. T. Press, Cambridge, Mass.
9. *Gilquin G., Gries S. Th.* (2009), Corpora and experimental methods: a state-of-the-art review, Corpus Linguistics and Linguistic Theory, Vol. 5(1), pp. 1–26.
10. *McEnery T., Wilson A.* (2001), Corpus Linguistics: An Introduction. University Press, Edinburgh.
11. *Paducheva E. V.* (1987), Particle *že*: semantics, syntax, prosody [La particule *že*: semantique, syntaxe et prosodie], Narrative particles in modern Russian [Les particules énonciatives en Russe contemporain], Vol. 3, pp. 11–44.
12. *Riemer N.* (2009), Grammaticality as evidence and as prediction in a Galilean linguistics, Language Sciences, Vol. 31, pp. 612–633.
13. *Sampson G.* (1992), Probabilistic parsing, Directions in Corpus Linguistics, Mouton de Gruyter, Berlin, pp. 425–447
14. *Schütze C. T.* (1996), The empirical base of linguistics: Grammaticality judgments and linguistics methodology, The University of Chicago, Chicago.
15. *Spencer A., Luis Ana R.* (2012), Clitics: an introduction, Cambridge Textbooks in Linguistics, Cambridge University Press, Cambridge.
16. *Tremblay A.* (2005), Theoretical and methodological perspectives on the use of grammaticality judgment tasks in linguistic theory, Second Language Studies, Vol. 24 (1), pp. 129–167.
17. *Wackernagel J.* (1892), On a law about Indo-European word order [Über ein Gesetz der indogermanischen Wortstellung], Indo-European Studies [Indogermanische Forschungen], Vol. 1, pp. 333–436.

References

1. *Bonno K., Kodzasov S. V.* (1998), Semantic variation of discourse words and its effect on the linearization and intonation (on the example of particles ‘že’ and ‘ved”) [Semanticheskoe var’irovanie diskursivnyh slov i ego vlijanie na linearizaciju i intonirovanie (na primere chastic «zhe» i «ved’»)], The discourse words of the Russian language: the experience of context-semantic description [Diskursivnye slova russkogo jazyka: opyt kontekstno-semanticheskogo opisanija], Metatekst, Moscow, pp. 382–446.
2. *Chomsky N.* (1957), Syntactic Structures, Mouton and Co., The Hague.

3. *Chomsky N.* (1962), Explanatory Models in Linguistics, Logic, Methodology, and Philosophy of Science, Stanford University Press, Stanford, Calif., pp. 528–550.
4. *Chomsky N.* (1965), Aspects of the Theory of Syntax, The M. I. T. Press, Cambridge, Mass.
5. *Gilquin G., Gries S. Th.* (2009), Corpora and experimental methods: a state-of-the-art review, Corpus Linguistics and Linguistic Theory, Vol. 5(1), pp. 1–26.
6. *Kiselev I. A.* (1976), Particles in modern East Slavic languages [Chasticy v sovremennykh vostochnoslavjanskikh jazykakh], BGU, Minsk.
7. *McEnery T., Wilson A.* (2001), Corpus Linguistics: An Introduction. University Press, Edinburgh.
8. *Paducheva E. V.* (1987), Particle *že*: semantics, syntax, prosody [La particule *že*: semantique, syntaxe et prosodie], Narrative particles in modern Russian [Les particules énonciatives en Russe contemporain], Vol. 3, pp. 11–44.
9. *Riemer N.* (2009), Grammaticality as evidence and as prediction in a Galilean linguistics, Language Sciences, Vol.31, pp. 612–633.
10. *Sampson G.* (1992), Probabilistic parsing, Directions in Corpus Linguistics, Mouton de Gruyter, Berlin, pp. 425–447
11. *Schütze C. T.* (1996), The empirical base of linguistics: Grammaticality judgments and linguistics methodology, The University of Chicago, Chicago.
12. *Spencer A., Luis Ana R.* (2012), Clitics: an introduction, Cambridge Textbooks in Linguistics, Cambridge University Press, Cambridge.
13. *Tremblay A.* (2005), Theoretical and methodological perspectives on the use of grammaticality judgment tasks in linguistic theory, Second Language Studies, Vol. 24 (1), pp. 129–167.
14. *Valova E. A.* (2014), Syntactic properties of Russian enclitic particle ‘zhe’ in the diachronic aspect: a corpus-based study [Sintaksicheskie svojstva enkliticheskoj chasticy zhe v diahronicheskom aspekte: korpusnoe issledovanie], Scientific and technical information. Series 2. Information processes and systems [Nauchno-tehnicheskaja informacija. Serija 2. Informacionnye processy i sistemy.], Vol. 10, pp. 31–36.
15. *Valova E. A.* (2014), Syntactic properties of Russian enclitic particle ‘že’ [Sintaksicheskie svojstva russkoj jenkliticheskoj chasticy zhe], Bulletin of an orthodox St. Tikhon humanitarian University. Series 3. Philology. [Vestnik Pravoslavnogo Svjato-Tihonovskogo gumanitarnogo universiteta. Serija 3. Filologija.], Vol. III:4 (39), pp. 16–33.
16. *Wackernagel J.* (1892), On a law about Indo-European word order [Über ein Gesetz der indogermanischen Wortstellung], Indo-European Studies [Indogermanische Forschungen], Vol. 1, pp. 333–436.
17. *Zaliznyak A. A.* (2008), Old Russian enclitics [Drevnerusskie enklitiki], Jazyki slavjanskih kul‘tur, Moscow.

«КАК ГОВОРИТСЯ, СТАТЬЯ ЕСТЬ СТАТЬЯ»: НЕКОТОРЫЕ АСПЕКТЫ ФУНКЦИОНИРОВАНИЯ ТАВТОЛОГИЙ В КОММУНИКАЦИИ[1]

Вилинбахова Е. Л. (e.vilinbakhova@spbu.ru)

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Ключевые слова: тавтологические конструкции, русский язык, микросинтаксис, семантика, прагматика

"AS THEY SAY, AN ARTICLE IS AN ARTICLE": SOME ASPECTS OF USE OF TAUTOLOGIES IN COMMUNICATION

Vilinbakhova E. L. (e.vilinbakhova@spbu.ru)

St. Petersburg State University, St Petersburg, Russia

In most studies dedicated to tautologies it goes without saying that these constructions are commonly used in everyday speech. Further analysis is based on the hearer's point of view concentrating mostly on possible ways of interpretation of tautologies. At the same time the perspective of the speaker remains largely unexplored. This study based on Internet and corpus data [RNC] deals with some aspects of use of tautologies in communication in order to understand why the speaker should opt for uttering tautologies instead of being more straightforward and what communicative profit he gets for that. It seems that advantages of using tautologies for the speaker are based on their structural and semantic features: (a) their recognizable form *X cop X* that makes tautologies look like a cliché; (b) their possibility to appeal to mutual knowledge; (c) the unquestionable truth of their literal meaning. First, when the speaker uses tautologies as clichés with expressions "as they say", *etc*., he makes his personal opinion look like a common wisdom of linguistic community. Next, when the speaker emphasizes that he appeals to mutual knowledge, he makes the hearer look as like-minded person, therefore the hearer's possible disagreement is regarded as a refusal of (expected) support and solidarity and requires more effort. Finally, the fact that the literal meaning of tautologies is undeniable helps the speaker escape of the responsibility of false implicature; defend his opinion using so-called *deep tautologies*; close the discussion whenever it is more convenient to him.

Key words: tautologies, Russian language, microsyntax, semantics, pragmatics

[1] Работа выполнена при поддержке гранта СПбГУ 31.42.1001.2016.

1. Вводные замечания

В большинстве работ, посвященных языковым тавтологиям вида *Война есть война* или *Любовь — это любовь*, отмечается, что данные конструкции, будучи неинформативными в буквальном смысле, тем не менее, активно используются говорящими и успешно интерпретируются слушателями (см., например, [Апресян 1995], [Булыгина, Шмелев 1997], [Падучева 2004], а также работы, приведённые ниже). Как правило, дальнейший анализ материала идёт именно с позиции слушателя: рассматривается, каким образом тавтологические высказывания получают интерпретацию и какие факторы помогают адресату правильно понять собеседника. В настоящее время можно выделить три подхода: прагматический (тавтологии универсальны и могут быть осмыслены благодаря общим прагматическим принципам), см. [Grice 1975], [Levinson 1983], [Ward, Hirschberg 1991], семантический (значение тавтологических паттернов конвенционально и может различаться даже в пределах одного языка) см. [Wierzbicka 1987, 1988, 1991], и семантико-прагматический (разные модели тавтологий могут иметь свои значения «по умолчанию», однако фоновые знания и контекст оказывают решающее воздействие на окончательную интерпретацию) см. [Fraser 1988], [Escandell Vidal 1990], [Miki 1996], [Autenrieth 1997], [Bulhof, Gimbel 2001], [Meibauer 2008], [Rhodes 2009], [Kwon 2014].

Позиция говорящего обычно остаётся за рамками исследований, как и вопросы, зачем автору высказывания использовать тавтологии в коммуникации, какие преимущества он при этом получает, и т. д., хотя в некоторых работах и упоминается, что «коммуникативная выгода» для говорящего, безусловно, присутствует, см. [Miki 1996], [Bulhof, Gimbel 2001].

В настоящей статье на материале Национального корпуса русского языка (далее — НКРЯ) и интернет-источников предпринята попытка описать некоторые аспекты функционирования тавтологий в коммуникации с позиции говорящего. Данный ракурс позволяет отнести тавтологии к «приёмам непрямого воздействия на слушающего», совокупность которых Т. В. Булыгина и А. Д. Шмелёв называют «языковой демагогией» [Булыгина, Шмелёв 1997: 461]. Её суть состоит в том, что «идеи, которые необходимо внушить слушающему, не высказываются прямо, а навязываются ему с помощью особых языковых механизмов» [Там же]: например, эмоционально окрашенных слов (ср. *сговор* vs. *соглашение, политикан* vs. *политик*) или конструкций.

Представляется, что способность тавтологий выступать в качестве подобных «приёмов» следует из их структурных и семантических характеристик. Во-первых, это особая форма конструкций *X сор X*, благодаря которой тавтологические высказывания могут быть поданы как клише (см. раздел 2); во-вторых, это их способность апеллировать к общим фоновым знаниям собеседников о предмете речи, (см. раздел 3); в-третьих, это их логически неоспоримая истинность в буквальном значении (см. раздел 4). В заключении представлены итоги и перспективы исследования, а также обозначен его прикладной потенциал.

2. Тавтологии как клише

Тавтологии, устанавливающие тождество[2], в русском языке имеют несколько синтаксических моделей, наиболее распространённые из которых: *X есть X* и *X — это X*. Формальные ограничения на заполнение переменных определяются референциальным статусом элементов конструкции: в метаязыковых тавтологиях, когда речь идёт о значении языковой единицы, используемое выражение может быть практически любым (см. [Вилинбахова 2015]); во всех прочих случаях преобладают именные группы.

Некоторые тавтологии в русском языке лексикализованы и воспринимаются носителями как «пословицы» с фиксированным значением, например, *Закон есть закон*[3] 'закон для всех одинаков', см. (1). Сюда же можно отнести обороты *Бизнес есть бизнес* 'на бизнес не должны влиять личные отношения, эмоции и пр.' и *Приказ есть приказ* 'приказ надо выполнять несмотря ни на что', см. (2), (3).

(1) А: ***Пословица Закон есть закон***. *помогите объяснить пожалуйста срочно заранее спасибо*
Б: *закон одинаков для всех богатых бедных крутых лохов и т. п. виноват сел будь ты хоть кто*[4]

(2) *Я полностью с Лолой согласна, но здесь дело в политике не процентов на 25 %, остальное только коммерция. Не даром у американцев есть* ***пословица «бизнес есть бизнес»***.

(3) *<Главный минер> посылал бегать, прыгать и подтягиваться вместо себя подчинённых. Это были мичман стопка и Журавлёв.* ***Как говорится, приказ есть приказ!*** [Ирина ТВЕРИТИНОВА. Хищения на Северном флоте: офицер украл миллион при помощи зачета… по физподготовке! // Комсомольская правда, 2012.12.07]

---

[2] В настоящей работе не рассматривались прочие разновидности тавтологий: условные, взаимоисключающие, подчинительные, относительные и сопоставительные, см. [Ward, Hirschberg 1991], [Autenrieth 1997], [Meibauer 2008].

[3] Закреплённость в языковой системе данного оборота отмечена ещё Ю. Д. Апресяном: «Устойчивые (лексикализованные) обороты типа *Закон есть закон* должны быть исключены из рассмотрения, поскольку они имеют единственное (словарное) толкование» [Апресян 1995: 166]. В английском языке А. Вежбицкая приводит в качестве устойчивых оборотов следующие: *Fair is fair* (букв.) 'Справедливо значит справедливо', *Enough is enough* 'Хватит значит хватит', *A deal is a deal* 'Уговор есть уговор', *Business is business* 'Бизнес есть бизнес' [Wierzbicka 1991: 404].

[4] Примеры взяты из Интернета, если не указано иное. Из соображений краткости адреса не указаны, но, разумеется, в материалах они присутствуют. Орфография и пунктуация оригинала почти полностью сохранены.

В примере (3) тавтология *Приказ есть приказ* сочетается с метатекстовым элементом *как говорится*, который, наряду с оборотами *как говорят, что называется* и т. д., указывает на употребительность и широкую эксплуатируемость выражения и используется с пословицами, модными словами, единицами бытовой афористики и пр., см. [Шмелёва 1987]. Однако, как оказалось, данные элементы сопровождают и «свободные» тавтологические высказывания.

(4) Наш проигрыш в хоккее, наше третье место, это, конечно, не совсем здорово, тем более, действительно не в кого кинуть камень, потому что играли наши ребята замечательно. Это, ***что называется «игра есть игра»,*** хотя конечно, мы претендовали на первое место, но нужно спокойно принимать удары. [Пропагандистская истерика Первого канала вокруг двойной «бронзы» России просто нелепа. Мнения экспертов // Новый регион 2, 2007.05.14].

(5) ***Путин есть Путин, как говорится****, наш Президент не умеет только одного — проигрывать. И тут турецкая коса нашла на камень.*

(6) А: *А вот теперь как раз и не верится, что всё так просто закончилось.*
Б: *Согласна, но,* ***как говорится, жизнь — это жизнь****. Фанфик мне очень понравился. Все описано просто шикарно!*

С одной стороны, можно было бы предположить, что устойчивость, на которую указывают обороты *как говорится* и др., относится только к форме, см. их допустимость с некоторыми нестандартными синтаксическими конструкциями с легко узнаваемой структурой и нетривиальным семантическим инвариантом, например, *Z X-овать не X-овал (но…P)* ‘Z не сделал X, но Z сделал P; P — нечто менее сильное, чем X’ [Иомдин 2013: 314]:

(7) ***Как говорится, видеть не видел****, но косвенных подтверждений дофига, а фантазию не остановить!*

С другой стороны, такие конструкции не сочетаются с оборотами *как известно, как все знают* и пр., см. (8), сопровождающие клише с ***общеизвестным, устойчивым содержанием***: пословицы, поговорки, цитаты и т. д., — и тавтологии, см. (9), хотя, за исключением оборотов в примерах (1)–(3), значение последних во многом определяется коммуникативным намерением говорящего[5].

(8) *?? **Как известно, видеть не видел**, но косвенных подтверждений дофига, а фантазию не остановить!*

---

[5] См. эксперимент на материале английского языка в [Gibbs, McCarrell 1990] на материале английского языка, где в разных контекстах одни и те же примеры тавтологий получали у испытуемых разную интерпретацию. В русском языке также можно найти примеры, когда модели *X есть X* и *X — это X* с различной семантикой (см. [Булыгина, Шмелёв 1997: 506–509]) оказываются взаимозаменимы или одна и та же модель передаёт разные значения.

(9) При консервировании все огурцы проходят множественный контроль, но, ***как известно, человеческий фактор есть человеческий фактор***[6].

Таким образом, тавтологии могут быть поданы как клише, что маркируется соответствующими безличными оборотами, которые «предполагают значение лица 'не я'; 'все, но не я' [Шмелёва 1987: цит. по эл. версии], и / или пунктуационно, т. е. говорящий частично снимает с себя ответственность сказанное, подчёркивая его широкую употребительность: 'все так говорят', 'все так считают'. Кроме того, он может использовать якобы устойчивое выражение как дополнительный аргумент для убеждения слушателя / аудитории, см. (10).

(10) *В-третьих, это, конечно же, комфорт и уют,* ***как говорится, дом есть дом,*** *и ничего казённого там нет.*

В данном примере автор, рекламируя агентство по предоставлению в аренду квартир посуточно, отстаивает спорную точку зрения, согласно которой в незнакомом городе лучше снимать квартиру, чем останавливаться в гостинице. Он приводит свой третий, последний аргумент после других доводов — о низкой стоимости и пр., и на месте тавтологии могла бы стоять пословица (*В доме и стены помогают, В гостях хорошо, а дома лучше,* и т. д.). Однако значение пословиц фиксировано, в них речь идёт в первую очередь о настоящем, своём доме, а не о съёмной квартире, что не очень подходит автору, поэтому в качестве аргумента привлекается тавтология, «не отягощённая» заранее заданным значением.

В целом, тавтологическая форма позволяет говорящему, с одной стороны, передать практически любое значение, а с другой стороны, подать высказывание как общеупотребительное клише с устойчивым содержанием, что придаёт ему дополнительный вес и убедительность[7].

3. Тавтологии как апелляция к общим фоновым знаниям

Важное свойство тавтологий, которое активно эксплуатируется в коммуникации, это их способность отсылать к общим фоновым знаниям о предмете речи. В литературе эту общую информацию обозначают по-разному: коннотации [Апресян 1995]; «мысленное досье референта» [Булыгина, Шмелёв 1997], ассоциации [Падучева 2004]; разделяемые убеждения (*shared beliefs*) [Miki 1996], стереотипы [Gibbs, McCarrell 1990], [Autenrieth 1997], [Meibauer 2008], и т. д.

---

[6] Также встретились примеры, где авторы выделяли тавтологии кавычками, см. (4), (21) и след.: *Мнения экспертов по поводу готовности города разделились: одни уверены, что* ***«Москва — это Москва»*** *и поэтому все пройдет отлично, а другие считают, что мундиаль сорвется из-за пробок.*

[7] По мнению М. В. Эсканделл Видаль, похожим образом функционируют риторические вопросы: «интересно то, что <риторические вопросы> используются, чтобы выдать за общепринятую точку зрения суждение, которое является не чем иным, как личным мнением говорящего (перевод мой — Е. В.)» [Escandell Vidal 2014: 189].

Допустимая степень известности информации зависит от семантического типа конструкций: для «классических» тавтологий с термовым референциальным статусом элементов подразумеваемая информация могут быть очень «закрытой» (знать о конкретных свойствах объекта могут только собеседники); для метаязыковых тавтологий «главное» значение слова должно быть известно всему языковому коллективу [Вилинбахова 2015: 605].

В любом случае, независимо от того, насколько общими или локальными являются фоновые знания, автор тавтологического высказывания даёт собеседнику понять следующее:

(а) говорящий рассчитывает, что слушатель разделяет с ним определённые знания о предмете речи;
(б) говорящий может не проговаривать эти сведения эксплицитно;
(в) слушатель способен сам вывести правильную инференцию и понять говорящего.

«Понять» в данном случае используется сразу в двух значениях: ‘успешно провести расшифровку языкового сигнала’ и ‘войти в положение говорящего и согласиться с ним’, см. примеры с выражениями *сам(и) понимаешь(-ете) / знаешь(-ете), мы оба / с тобой / все понимаем / знаем,* и др. которые регулярно сопровождают тавтологии:

(11) *Ладно, Гриша, — вздохнул Антон. — Я сейчас вызываю группу, поедем на место, покажешь гараж. А тебя потом в камеру.* ***Сам понимаешь, порядок есть порядок****. — Да не вопрос, начальник, — осклабился Дубинюк. — мы всё понимаем. Командир в беде не бросит.* [Александра Маринина. Последний рассвет (2013)]

(12) [Майор, муж] *извините / товарищ генерал / что пришлось побеспокоить… Ну /* ***сами понимаете / работа … есть работа****.* [Павел Чухрай. Водитель для Веры, к/ф (2004)]

(13) *Поправь меня, если я ошибаюсь.* ***Мы все понимаем: кризис — это кризис****. Дефолт и пирамида ГКО* [Виктор Левашов. Заговор патриота (2000)]

(14) *Хочу на праздники мужа за ней <коляской> отправить, но,* ***сами понимаете, мужчина есть мужчина****.*

Таким образом, говорящий показывает, что по умолчанию воспринимает собеседника как «своего человека», разделяющего с ним общие убеждения и ценности, и потому рассчитывает на его понимание, сочувствие и т. д.

Следует отметить, что указанный процесс происходит в одностороннем порядке: адресат может в действительности иметь другое мнение и не считать, например, что мужчины ничего не понимают в детских вещах, см. (14). Ивсёжедляслушателявыйтиизнавязанной говорящимролиединомышленника становится сложнее и требует больше усилий, чем возражение эксплицитному

утверждению: речь идёт уже не о разногласиях по поводу частного вопроса, а об отказе быть для говорящего «своим».

Между тем, зачастую такое «союзничество» оказывается собеседнику совершенно не выгодно, см. (11), (12) и примеры ниже, когда тавтологии служат оправданием говорящему в неприятной для слушателя ситуации.

(15) *— Моя мама… — напомнил Анакин.*
— А да, Шми. Она больше не моя. Я ее продал.
— Продал?
Падме успела поймать его за руку.
— Несколько лет назад, — пояснил Уотто, с гудением кружа над табуретом. — ***Мне жаль,*** *Ани,* ***но сам знаешь, дело есть дело****.*

Также одностороннее сокращение дистанции даёт говорящему больше свободы, чем это было бы допустимо для постороннего, например, позволяет ему задавать личные вопросы, см. (16).

(16) ***Вопрос, наверное, бестактный, но жизнь есть жизнь****. Могли бы вы представить, что место Раисы Максимовны в вашей жизни сегодня может занять другая женщина?* [Ольга ВАНДЫШЕВА. Михаил Горбачев: Жениться я не собираюсь // Комсомольская правда, 2001.03.02]

В целом, говорящий, используя тавтологии как апелляцию к общим знаниям, даёт понять, что воспринимает собеседника как своего единомышленника и ожидает от него солидарности и поддержки. Такая стратегия, с одной стороны, позволяет говорящему оправдывать свои п(р)оступки, а с другой, снижает риск нежелательной реакции адресата.

4. Тавтологии как неоспоримая истина

Во многих работах отмечается, что тавтологии чаще всего используются в переносном смысле, а их буквальное значение — утверждение тождества объекта самому себе — описывается как неинформативное и не учитывается при дальнейшем анализе, см., например, [Fraser 1988], [Autenrieth 1997], [Meibauer 2008] и др. Между тем, само наличие буквального значения и его формальная истинность также могут эксплуатироваться говорящим и оказывать влияние на слушателя.

Во-первых, в многочисленных ситуациях, когда тавтология имеет переносный смысл, и собеседнику приходится самому выводить нужную инференцию, говорящий при необходимости может избежать ответственности за своё суждение, см. замечание И. М. Кобозевой о косвенных речевых актах (к которым, в данном случае, относятся и тавтологические высказывания): «косвенность «развязывает руки» автору высказывания, позволяя ему, если это понадобится, сказать, что он имел в виду только буквальный смысл сказанного» [Кобозева 2003: цит. по эл. версии].

Так в примере (17) у участников коммуникации есть общее «мысленное досье» войны, поэтому ожидаемые импликатуры говорящего могут быть следующими: ‘случиться могло всё что угодно’, ‘возможно, она погибла’ и т. п. Тем не менее, на уточняющий вопрос, например, «Вы думаете, что её убили?» автор тавтологического высказывания всегда может ответить «Я этого не говорил».

(17) — Как вы думаете, что могло случиться с Цзинь Фын? — ***Война есть война***, — ответил начальник разведки и, направив свет фонаря на новое препятствие, предупредил: — Пожалуйста, не споткнитесь. [Н. Н. Шпанов. Связная Цзинь Фын (1935–1950)]

Следует подчеркнуть, что тавтологические высказывания формально всегда остаются истинными, даже если несут ложную импликатуру. Таким образом, в случае, когда обсуждаемый субъект X в конкретной ситуации поступил иначе, чем можно было бы от него ожидать, высказывание *X есть X* как ответ на вопрос, затрагивающий его поведение, реакции и пр., скорее всего, введёт собеседника в заблуждение, при этом не являясь ложью в полном смысле этого слова[8].

Во-вторых, существует особый подкласс ***глубинных тавтологий*** (англ. *deep tautologies*), впервые описанных Й. Булхофом и С. Гимбелом [Bulhof, Gimbel 2001], которые предполагают именно буквальную интерпретацию, т. е. выражают значение тождества объекта самому себе. Авторы указывают на два условия *неотчуждаемости* и *равноценности*, выполнение которых (одного или обоих сразу) подразумевает говорящий, используя в речи глубинные тавтологии, и формулируют их следующим образом:

1) Условие неотчуждаемости: если объект отвечает условиям, достаточным для того, чтобы быть сущностью x, то никакие дополнительные свойства не могут исключить его из категории X;
2) Условие равноценности: если объект является сущностью x, то он не может быть x в большей или меньшей степени» [Bulhof, Gimbel 2001: 287].

Применяя глубинные тавтологии, говорящий апеллирует к формальному пониманию языкового выражения и намеренно игнорирует обстоятельства, которые с точки зрения морали, здравого смысла, культурных представлений и пр. следовало бы учитывать[9].

---

[8] Например, в следующем сконструированном на основе социальной рекламы диалоге: — Как новый рабочий справился с заданием? / — *Мигрант есть мигрант*, ответная реплика не позволяет понять, что рабочий-мигрант выполнил работу очень хорошо, и склоняет к противоположному выводу, оставаясь формально истинной.

[9] Кажется, метаязыковые тавтологии (см. [Вилинбахова 2015]) можно было бы считать подмножеством глубинных тавтологий с подразумеваемым условием неотчуждаемости, где «дополнительные свойства» языковой сущности определяются контекстом. Скажем, в примере «*Хочу есть» значит «хочу есть», а не своди меня в ресторан* фраза «хочу есть», имеет дополнительные условия ‘обращённая к молодому человеку и произнесённая девушкой’, поэтому теоретически может означать пожелание пойти в ресторан. Впрочем, в данной работе этот вопрос остаётся за рамками исследования.

(18) *Что касается вопроса, **измена есть измена**, хоть душой хоть телом.*

(19) ***Убийство — это убийство**, даже если оно совершено в ходе самообороны.*

В примере (18) говорящий утверждает, что возможное общение мужа с другой женщиной (в данном случае, речь идёт о скайп-сообщениях) эквивалентно физической измене, хотя возможна и альтернативная позиция. В примере (19) интуитивно кажется, что убийство в целях самообороны должно квалифицироваться иначе, чем прототипическое намеренное и подготовленное убийство, но автор высказывания это отрицает[10]. Важно, что в привёдённых текстах обе противоречащие друг другу точки зрения имеют право на существование, однако высказывания с глубинными тавтологиями сложнее оспорить в силу их формальной истинности.

Логическая справедливость тавтологий в буквальном смысле, независимо от наличия у них переносного значения, объясняет ещё одно их свойство — выступать маркером закрытия темы. Эта характеристика отмечается в [Levinson 1983] и [Wierzbicka 1991], и подтверждается реальными примерами из НКРЯ и интернета.

(20) А: *Скажите, а как совместить консерватизм растущий и в обществе, и в деятельности вашей партии, и в детальности Путина с тем, что ваша личная жизнь, я вас поздравляю с законным браком, она не соответствует консерватизму. У вас не первый брак, скажем так.*
Б: *Давайте мы скажем себе, что **личная жизнь есть личная жизнь**, и на этом поставим точку.*

(21) *Он и правда собирался за один год два класса проскочить — не разрешили. Он в министерство гонял, там тоже оказались консерваторы. И Эфэф за него хлопотал, ничего не помогло. Ему сказали, что «**закон есть закон**», и точка.* [Владимир Железников. Каждый мечтает о собаке (1966)]

В примере (20) государственный деятель при помощи тавтологии прекращает разговор на неудобную для себя тему; в примере (21) чиновники из министерства отказывают школьнику в просьбе, также используя тавтологию. В обоих случаях продолжение дискуссии для участников коммуникации становится затруднительным, превращаясь в бессмысленный спор с тривиальным утверждением.

Таким образом, логически неоспоримая истинность тавтологий в буквальном значении даёт говорящему преимущество в коммуникации: при передаче имплицитного содержания это «страховка» на случай, если почему-либо придётся отказываться от своего суждения; при использовании глубинных

---

[10] Вспоминается также рассказ Р. Брэдбери «Наказание без преступления» (1950), где герой убивает куклу — копию своей жены, и за это его приговаривают к смертной казни; логика судей может быть выражена глубинной тавтологией *убийство есть убийство*, несмотря на то, что герой не хотел убивать свою жену в реальности и именно для освобождения от негативных эмоций обратился в специальное агентство.

тавтологий это презентация своей точки зрения как единственно верной, и, наконец, при нежелательном повороте в разговоре это возможность закончить его или сменить тему.

5. Заключение

В настоящей статье была предпринята попытка рассмотреть некоторые аспекты функционирования тавтологий в коммуникации, чтобы показать, зачем говорящему использовать данные конструкции в речи и какие преимущества он при этом получает. Было установлено, во-первых, что тавтологические высказывания могут быть поданы как общеизвестные клише, передавая при этом личное мнение их автора. Во-вторых, говорящий, используя тавтологии как отсылку к общим фоновым знаниям, в одностороннем порядке навязывает собеседнику роль своего единомышленника. В-третьих, очевидная истинность конструкций в буквальном значении даёт возможность автору высказывания при необходимости апеллировать именно к этому «верному по определению» значению, отрицая возможные импликатуры; представить свою точку зрения как единственно правильную; положить конец нежелательной дискуссии. В свою очередь, адресат оказывается в ситуации, когда несогласие с говорящим может представить его в неприглядном свете, будучи воспринято либо как отрицание общеизвестной прописной истины, либо как отказ в ожидаемой солидарности и поддержке, либо как спор с логически неопровержимым суждением.

Говоря о перспективах исследования, представляется интересным проанализировать в дальнейшем «отрицательные» тавтологии вида *X neg cop X* (см. [Булыгина, Шмелёв 1997], [Meibauer 2008]), для которых одни характеристики, — например, отсылка к общим фоновым знаниям — сохраняются, а другие меняются на противоположные: вместо «неоспоримой истины» в буквальном значении мы имеем парадокс, противоречие.

Обращаясь к прикладному потенциалу исследования[11], стоит отметить, что изучение импликатур становится все более важным для таких задач в области автоматической обработки текста, как RTE (Recognizing Textual Entailment, возможный русскоязычный термин — *логический вывод по фрагменту текста*), когда проверяется (не)соответствие гипотезы H определённой информации, содержащейся во фрагменте текста T (см., например, [Bos, Markert 2005]). Описанные «коммуникативные» свойства тавтологий могут быть полезны для выявления скрытых намерений авторов и повышения вероятности выведения правильных инференций для соответствующих конструкций, тем самым внося свой вклад в общую интерпретацию текста.

---

[11] Автор благодарен анонимному рецензенту за указание на возможность практического применения работы в компьютерной лингвистике.

Литература

1. *Апресян Ю. Д.* (1995), Коннотации как часть прагматики слова, Избранные труды. Т. II. Интегральное описание языка и системная лексикография, «Языки русской культуры», Москва, с. 156–178.
2. *Булыгина Т. В., Шмелёв А. Д.* (1997), Языковая концептуализация мира (на материале русской грамматики), Школа «Языки русской культуры», Москва.
3. *Вилинбахова Е. Л.* (2015), Статья значит статья: об одном классе тавтологических конструкций в русском языке, Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог», Изд-во РГГУ, Москва, вып. 14 (21): в 2 т., т. 1, с. 638–649.
4. *Иомдин Л. Л.* (2013), Читать не читал, но…: об одной русской конструкции с повторяющимися словесными элементами, Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог», Изд-во РГГУ, Москва, вып. 12 (19): в 2 т., т. 1, с. 309–322.
5. *Кобозева И. М.* (2003), Лингво-прагматические аспекты анализа языка СМИ, Язык СМИ как объект междисциплинарного исследования: учеб. пособие, МГУ, Москва, режим доступа эл. версии http://evartist.narod.ru/text12/08.htm
6. *Национальный корпус русского языка, режим доступа: www.ruscorpora.ru*
7. *Падучева Е. В.* (2004), Динамические модели в семантике лексики, «Языки славянской культуры», Москва.
8. *Шмелева Т. В.* (1987), «Так сказать» и «как говорится», Служебные слова: сборник статей, Новосибирск, с. 125–132, режим доступа электронной версии: http://www.novsu.ru/file/1077482
9. *Autenrieth T.* (1997), Tautologien sind Tautologien, Pragmatik. Implikaturen und Sprechakt, Westdeutscher Verlag, Opladen, pp. 12–32.
10. *Bos J., Markert K.* (2005), Recognizing textual entailment with logical inference, Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP), Vancouver, pp. 628–635.
11. *Bulhof J., Gimbel S.* (2001), Deep tautologies, Pragmatics and Cognition, Vol. 9–2, pp. 279–291.
12. *Escandell Vidal M. V.* (1990), Nominal tautologies in Spanish, доклад на Международной конференции по прагматике, Барселона.
13. *Escandell Vidal M. V.* (2014), Introducción a la pragmática, Ariel Letras, Barcelona.
14. *Fraser B.* (1988), Motor oil is motor oil: An account of English nominal tautologies, Journal of Pragmatics, Vol. 12, pp. 215–220.
15. *Gibbs R. W., McCarrell N. S.* (1990), Why boys will be boys and girls will be girls: understanding colloquial tautologies, Journal of Psycholinguistic Research, Vol. 19, pp. 125–145.
16. *Grice H. P.* (1975), Logic and conversation, in Syntax and Semantics, Vol. 3: Speech Acts, Academic Press, New York, pp. 41–58.
17. *Kwon I.* (2014), Categorization and its embodiment: Korean tautological constructions in mental spaces theory, Language Sciences, Vol. 45, pp. 44–55.

18. *Levinson S.* (1983), Pragmatics, Cambridge University Press, Cambridge.
19. *Meibauer J.* (2008), Tautology as presumptive meaning, Pragmatics and Cognition, Vol. 16, pp. 439–470.
20. *Miki E.* (1996), Evocation and tautologies, Journal of Pragmatics, Vol. 25, pp. 635–648.
21. *Rhodes R.* (2009), A Cross-linguistic comparison of tautological constructions with special focus on English, available at: www.linguistics.berkeley.edu/~russellrhodes/pdfs/taut_qp.pdf
22. *Ward, G. L., Hirschberg J.* (1991), A pragmatic analysis of tautological utterances, Journal of Pragmatics, Vol. 15, pp. 507–520.
23. *Wierzbicka A.* (1987), Boys will be boys: ‹Radical semantics› vs. ‘Radical pragmatics’, Language, Vol. 63, pp. 95–114.
24. *Wierzbicka A.* (1988), Boys will be boys: A rejoinder to Bruce Fraser, Journal of Pragmatics, Vol. 12, pp. 221–224.
25. *Wierzbicka A.* (1991), Cross-cultural pragmatics: the semantics of human interaction, Mouton de Gruyter, Berlin; New York.

References

1. *Apresyan Ju. D.* (1995), Connotations as a part of pragmatic meaning of a word [Konnotatsii kak chast’ pragmatiki slova], in Selected works: In 2 vol. Vol. II: Integrated description of language and systematic lexicograpgy [Izbrannye trudy. Vol. 2: Integral’noe opisanie yazyka i sistemnaya leksikografiya], Yazyki Russkoy Kul’tury, Moscow, pp. 156–178.
2. *Autenrieth T.* (1997), Tautologies are Tautologies [Tautologien sind Tautologien], in Pragmatics, Implicatures and Speech Acts [Pragmatik, Implikaturen und Sprechakte], Westdeutscher Verlag, Opladen, pp. 12–32.
3. *Bos J., Markert K.* (2005), Recognizing textual entailment with logical inference, Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP), Vancouver, pp. 628–635.
4. *Bulhof J., Gimbel S.* (2001), Deep tautologies, Pragmatics and Cognition, Vol. 9–2, pp. 279–291.
5. *Bulygina T. V., Shmelev A. D.* (1997), Linguistic conceptualization of the world (on the material of the Russian grammar) [Yazykovaya kontseptualizatsiya mira (na materiale russkoy grammatiki)], Shkola “Jazyki russkoy kul’tury”, Moscow.
6. *Escandell Vidal M. V.* (1990), Nominal tautologies in Spanish, talk given at the International conference of pragmatics (IPRA), Barcelona.
7. *Escandell Vidal M. V.* (2014), Introduction to pragmatics [Introducción a la pragmática], Ariel Letras, Barcelona.
8. *Fraser B.* (1988), Motor oil is motor oil: An account of English nominal tautologies, Journal of Pragmatics, Vol. 12, pp. 215–220.
9. *Gibbs R. W., McCarrell N. S.* (1990), Why boys will be boys and girls will be girls: understanding colloquial tautologies, Journal of Psycholinguistic Research, Vol. 19, pp. 125–145.

10. *Grice H. P.* (1975), Logic and conversation, in Syntax and Semantics, Vol. 3: Speech Acts, Academic Press, New York, pp. 41–58.
11. *Iomdil L. L.* (2013), Chitat' ne chital, no…: on a Russian Construction with Repeated Lexical Elements. [Chitat' ne chital, no…: ob odnoy russkoy konstruktsii s povtoryayushchimisya slovesnymi elementami], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2013" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2013"], Moscow, pp. 309–322.
12. *Levinson S.* (1983), Pragmatics, Cambridge University Press, Cambridge.
13. *Kobozeva I. M.,* (2003), Linguopragmatic aspects of analysis of the language of media [Lingvo-pragmaticheskie aspekty analiza yazyka SMI], The language of media as an object of interdisciplinary studies: textbook [Yazyk SMI kak ob"ekt mezhdistsiplinarnogo issledovaniya: uchebnoe posobie], MGU, Moscow, available at: http://evartist.narod.ru/text12/08.htm
14. *Kwon I.* (2014), Categorization and its embodiment: Korean tautological constructions in mental spaces theory, Language Sciences, Vol. 45, pp. 44–55.
15. *Meibauer J.* (2008), Tautology as presumptive meaning, Pragmatics and Cognition, Vol. 16, pp. 439–470.
16. *Miki E.* (1996), Evocation and tautologies, Journal of Pragmatics, Vol. 25, pp. 635–648.
17. *Paducheva E. V.* (2004) Dynamic models in lexical semantics [Dinamicheskie modeli v semantike leksiki], Jazyki slavyanskoy kul'tury, Moscow.
18. *Rhodes R.* (2009), A Cross-linguistic comparison of tautological constructions with special focus on English, available at: www.linguistics.berkeley.edu/~russellrhodes/pdfs/taut_qp.pdf
19. *Russian National Corpus* [Natsional'nyy korpus russkogo yazyka], available at: www.ruscorpora.ru
20. *Shmeleva T. V.* (1987), "So to say" and "as they say" ["Tak skazat'" i "kak govoritsa"], Auxiliary words [Sluzhebnye slova], Novosibirsk, pp. 125–132, available at: http://www.novsu.ru/file/1077482
21. *Vilinbakhova E. L.* (2015), Article means article: on one pattern of tautologies in Russian [Stat'ya znachit stat'ya: ob odnom klasse tavtologicheskih konstruktsiy v russkom yazyke], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2015" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2015"], Moscow, pp. 638–649.
22. *Ward, G. L., Hirschberg J.* (1991), A pragmatic analysis of tautological utterances, Journal of Pragmatics, Vol. 15, pp. 507–520.
23. *Wierzbicka A.* (1987), Boys will be boys: ‹Radical semantics› vs. ‹Radical pragmatics›, Language, Vol. 63, № 1, pp. 95–114.
24. *Wierzbicka A.* (1988), Boys will be boys: A rejoinder to Bruce Fraser, Journal of Pragmatics, Vol. 12, pp. 221–224.
25. *Wierzbicka A.* (1991), Cross-cultural pragmatics: the semantics of human interaction, Mouton de Gruyter, Berlin; New York.

THE ROLE AND APPLICATIONS OF EXPERT ERROR ANNOTATION IN A CORPUS OF ENGLISH LEARNER TEXTS

Vinogradova O. I. (olgavinogr@gmail.com)
Research University Higher School of Economics, Moscow, Russia[1]

The paper presents the rationale for the decisions that were taken in the set-up and further development of a learner corpus of student texts written in English by Russian learners of English, the only Russian learner corpus in the open access. The tool of manual expert annotation is in the focus of the present observations, and after introducing categorization of errors applied in annotation, the complicated cases that arose in annotation practices have been looked into followed by comparison of the annotation statistics over the three stages in the corpus development. For that purpose, texts annotated by different groups of participants in the process of two experiments were used to spot the problematic areas in annotation. The main pedagogical applications of the learner corpus in teaching EFL—the opportunities to create automated training exercises and placement and progress tests custom-made for specific groups of students—are outlined in the concluding part of the paper.

Key words: learner corpora; annotation; corpus research; computational tools

ЗНАЧЕНИЕ И ПРИМЕНЕНИЕ ЭКСПЕРТНОЙ АННОТАЦИИ ОШИБОК В КОРПУСЕ АНГЛОЯЗЫЧНЫХ УЧЕБНЫХ ТЕКСТОВ

Виноградова О. И. (olgavinogr@gmail.com)
НИУ ВШЭ, Москва, Россия

Ключевые слова: учебные корпуса; аннотирование; корпусные исследования; компьютерные инструменты

A learner corpus is a systematic computerised collection of texts that are written and/or oral productions of language learners. As all other corpora, a learner corpus is usually provided with convenient means of browsing and search options, with a system for marking the texts for pedagogical and/or research purposes, and ideally with additional visualisation of statistical processing of the search results. The first

[1] The study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) in 2016, and the author is a member of the team that has won a Research Team Project Competition in 2016 (16-05-0057 at https://www.hse.ru/en/science/scifund/nug).

researches in this area of computational linguistics date back to late 80s—early 90s, and the main achievements in the area have been well reviewed in the collection edited by Granger, Gilquin and Meunier 2013. The main point of interest for linguists working on the development of a learner corpus is the choice of annotation. Learner corpora are usually *error-tagged*, which means that spelling, lexical, and grammatical errors in the texts have been outlined with the help of a standardised system of error tags. The exhaustive list of important references for the discussion of the use of annotation in learner corpora can be seen in Pustejovsky and Stubbs 2013 and Wilcock 2009. The following researchers wrote on approaches to annotation in different learner corpora: Granger 2003, Hovy & Lavid 2010, and on the decisions concerning the choice of annotation systems, see Shtindlova et al. 2014, Lee et al. 2014, and many others.

This article provides rationale for the decisions on corpus annotation taken in setting up one of the first Russian learner corpora and to our knowledge the only learner corpus of English student texts in the open access: this corpus is free to search in and freely downloadable. The name of the corpus, REALEC, stands for Russian Error-Annotated English Learner Corpus, and its texts are now available at http://realec.org and at http://realec.org/hse/#/data_4_staff. The focus will be placed on evaluating how the chosen tools and the annotation workflow affect the results of annotation. The paper concludes by discussing the prospects of how manual expert tagging in this particular corpus can be used in creating a few pedagogical and research applications.

The corpus now comprises almost 3,400 pieces of students' writing (with about 838,000 word tokens), of which essays written in preparation for IELTS and during the examination of IELTS type make up the main part. It was initially set up as a pedagogical tool for EFL instructors who teach a course of general English, which includes preparation for IELTS, and also for professors teaching Academic Writing in English. The initial goals were to provide those instructors with the tool for marking written works submitted by their students, as well as to give instructors the opportunities to carry out their independent research, and at the same time to provide students with the easy means to see which errors prevail in their writing. To satisfy these three areas of need, expert error annotation was designed on the BRAT platform[2] (see Hovy (2015), p. 5 on growing popularity of BRAT).

At the present time REALEC has a well developed system of hierarchical tags to mark the errors, and these tags are shown above the text as labels in different colours along with suggestions on how to correct the error. REALEC error annotation scheme consists of four layers: error type, error cause, linguistic 'damage' caused by the error, and the impact of the error on general understanding of the text. The first of the annotation layers is the main source of knowledge about the mistake a particular student has made, so the paper only deals with this layer of annotation process, and the term *'annotation'* will be reserved in this paper for assigning tags that specify error type. The scheme includes 151 categories organised into a tree-like structure presented in Figure 1.

[2] Stenetorp et al. 2012

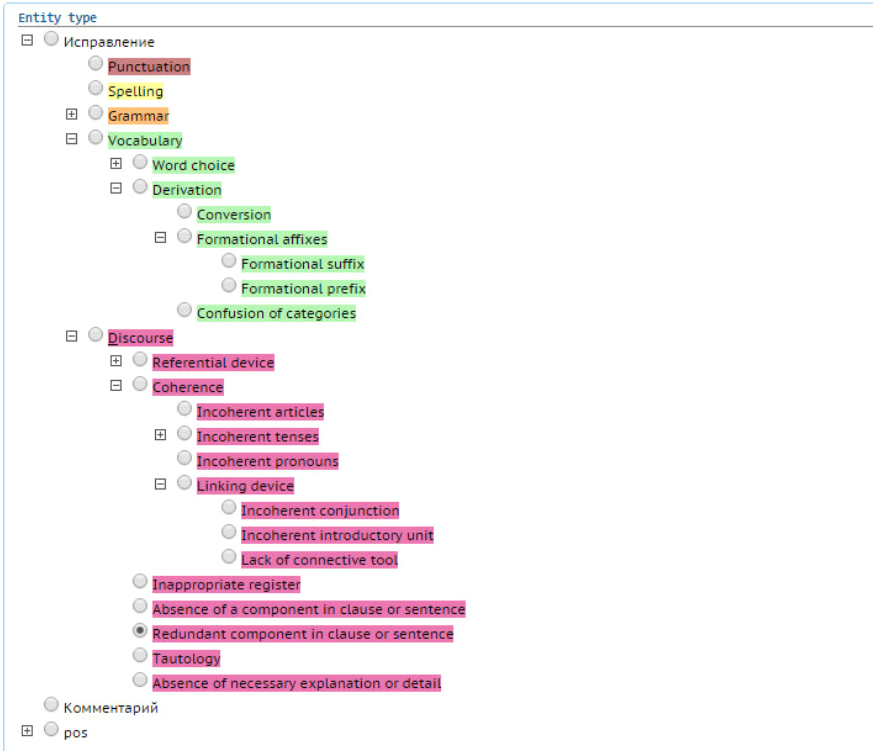


Fig. 1. Outline of the error-tagging scheme in REALEC[3]

In REALEC annotation, the following two important principles are observed: first, annotators mark as error spans only the areas of the text with clearly identified mistakes, and, second, they choose the most specific tag available in the scheme for the error they have spotted, with the exception of some special cases, when they can assign one of the more general tags.

Expert annotation in a learner corpus has to be continuously evaluated. On the question of the comparative evaluation of expert manual annotation and automated error annotation, there are three important points:

- Learner corpora are usually proprietary and often cannot be shared (Chodorow et al. 2010 and Chodorow et al. 2014). On the contrary, REALEC, as was mentioned above, is open for research.
- Learner corpora are as a rule expensive to annotate manually, and any alternative to time-consuming expert annotation has to be applied and tested. Some industrial applications have been reviewed by Chodorow et al. 2010 and Sorokin &

[3] As can be seen from a "+" in the little window, some error tags—namely, all of the *Grammar* area, *Word Choice* in *Vocabulary* area, and two tags in the *Discourse* area—*Referential device* and *Incoherent tenses*—are further subdivided into classes and subclasses of specific tags not demonstrated in Figure 1.

Forsyth 2008, but at present they do not seem to give a valid alternative to manual pedagogical endeavours.

- The last two decades have seen an explosion in the development of NLP tools that aim to detect and correct errors made by learners of English as a Second Language (ESL) or English as a Foreign Language (EFL), so to meet this growing need, annotation schemes have to be built into the approach that combines automated detection of simpler errors with expert annotation of more sophisticated ones. An approximation to such a system can, for instance, be seen in the CEA—computer-aided error analysis (Diez-Negrillo & Fernandez-Dominguez 2006); it is also presented in Yannakoudakis 2013. In the long run, there will arise the possibility of building a system that models human behaviour in the process of reading and making judgments about the value of someone's writing.
- When learner corpora are to be used to investigate learning process, high-quality corpus annotations as a basis for analyses are of great importance, and primarily it implies minimising the number of annotation errors for each annotator (Glaznieks et al. 2014). That is why annotation schemes are always subject to scrutiny in the process of using a learner corpus, and as an example of this, Bayerl 2008 illustrates, on the one hand, various forms of 'annotator drift' as annotators get tired over time, and on the other, how their mutual agreement levels change over time during work with the corpus. This is precisely the area of the current research interests.

To check how the level of precision of manual annotation affects annotator agreement, we looked into the results of an annotator agreement experiment carried out in REALEC in 2015. There were 10 annotators involved—one leading the experiment, three English instructors familiar with the annotation practices, another three without any exposure to the annotation process in REALEC, and three more—students in computer linguistics proficient in English. All the participants were instructed on tagging practices at the beginning of the experiment and were given 30 student essays 150–300 words each with error spans outlined by the leader of the experiment, so that they had to identify the error in the outlined areas and to look for the appropriate tag, or take off the mark if they did not see any mistake. The results—thirty texts tagged by ten annotators—were then subjected to two stages of research.

The first stage dealt with the procedure of calculating inter-rater agreement. The standard procedure is to use Krippendorff's alpha (further KA) (Krippendorff 2007; Hayes and Krippendorff 2007, Krippendorff 2012), Cohen's kappa or Fleiss's kappa. The goal of achieving a decent agreement among human annotators is difficult even for such an algorithm-prone system as specific grammar errors (see, for example, Bryant & Hwee Tu Ng 2015). Full agreement is almost never possible with any non-trivial annotation task, but the extent of agreement is still an important index of how reliable the adopted annotation method is.

Our 2015 experiment to check the rate of agreement among annotators was reported at the 8th International Corpus Linguistics Conference in Lancaster (Kutuzov, Kuznenko, Vinogradova 2015), so I will only briefly state the results here. There were the total of 2128 error category assignments involved. A topical question was how

to apply KA in view of the hierarchical nature of our annotation scheme, and we did it by transforming our nominal scale of tags into an interval scale. To explain, grammar errors differ one from another, but they are even more different from discourse errors. We assigned digital representations, or 'coefficients', to our error categories according to our intuitive knowledge of which categories are closer, so that tags belonging to closely related categories were assigned closer values. For the five macro-categories in REALEC, we assigned specific digital representations to subcategories. For example, the morphological part of macro-category Grammar is further divided into POS subcategories of Verb, Noun, etc. These tags are assigned different digital representations ("1", "4", "7", etc), whereas tags deeper down the hierarchy are assigned the same values as the upper ones. Between macro-categories we made 'gaps' 50 points wide. At the next level of the annotation scheme, we went down to the third-level subcategories (for example, Tense, Voice, Modals, etc). The same principle gave us the way to compute Krippendorff's alpha as if annotators had assigned interval digital values, and not nominal tags. As a result, we got Krippendorff's alpha = 0.57 for the second level annotation (tags like Noun, Verb, Word choice, Tautology, etc), even higher than at the upper level. The third level annotation had agreement rate equal to 0.55. Computing KA for the second and the third annotation levels as nominal categories (binary distance) gave only 0.5 and 0.4 correspondingly. The resulting index was satisfactory (KA = 0.57).

At the second stage, which has not been presented in a report or paper yet, the texts annotated in the experiment were used to research the cases of, and spot the reasons for, the lack of annotators' agreement. I compared the results of each participant in each of the three groups of annotators with the results of all participants from two other groups, and then calculated the average values for each three participants of groups of the type "EFL instructor familiar with annotation/English student or instructor unfamiliar with annotation/computer linguist". Fig. 2 shows the statistics for the average group.

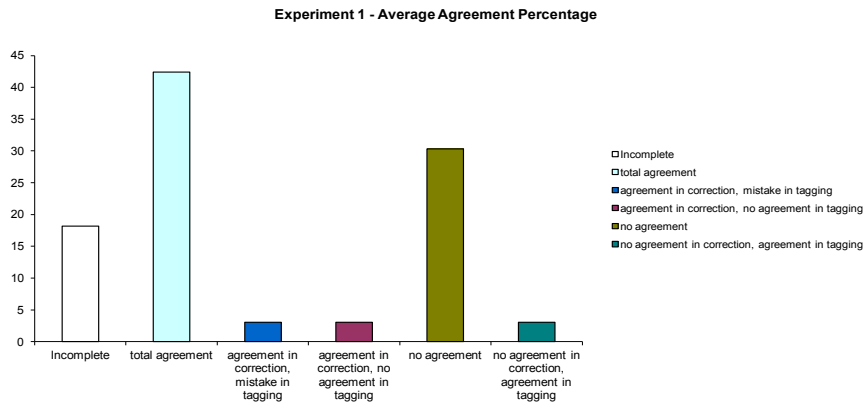


Fig. 2. Variation of agreement figures (average percentage) in agreement experiment

The source data for this graph represents the average figures shown in this experiment, namely, 33 error spans per text on average initially outlined, of which in 6 on average annotators did not find a mistake and thus did not assign any tags, and among the 27 error spans where some tags were assigned annotators agree on average over 17 errors and disagree over 10. The extent of agreement in this case can be different—annotators can agree about both the tag and the correction, or about one of them only. The following three examples illustrate it.

(1) *twice lucky> twice as lucky* (text 11) the same correction, different tags:
"Absence of certain component" (a vocabulary tag)—1 annotator
"Numerical comparison"—2 annotators
"Comparative degree of adverbs"—2 annotators—wrong tag!
"Prepositions"—1 annotator—wrong tag!
"Absense of a component in clause or sentence" (a discourse tag)—1 annotator

(2) —*twice lucky> double lucky* (text 11) different corrections, different tags ("Vocabulary"—1 annotator)

(3) *And there was the same situation in 2001 with only a few variations in five cities* (text 3) the same tags, different corrections: all annotators used a tag "Standard word order" and some discourse tag to change *cities* for *provinces*, as well as the tag "Preposition" to change *in* for *for* or *among,* and they used one more discourse tag—"Coherence"—to show the need for a change in the construction. Nevertheless, the resulting corrections were different:

>The same situation was in 2001, only with a few variations in five provinces
>And the situation was the same in 2001 with only a few differences for some provinces
>The same situation was in 2001 with only a few variations among five provinces
>The same situation was in 2001, only there were a few variations for the five provinces

In 2016, our goal was to trace the effect of changes that have taken place in our work over three years of active annotation practices. For this purpose, we collected data on the use of annotation tags in the following three areas of REALEC:

1. the initial student texts (essays, paragraphs, texts written in Academic Writing course, and theses) collected over the first year of using the corpus and tagged by a group of students—participants of the research seminar (below referred to as ESL); the total of 1,239 texts with 361,240 tokens of error annotation;
2. IELTS-type essays from different departments of the Higher School of Economics dating back to 2014–2015 academic year and annotated by students in the Bachelor's course in linguistics at the HSE as their summer practical work (below referred to as IELTS); the total of 1,941 texts with 433,523 tokens of error annotation;
3. essays written in preparation for IELTS-type examination by students of one EFL instructor and annotated by students themselves or in peer tagging under the supervision of their instructor (below referred to as *current subcorpus*

and labeled as 2ndYear 2015–2016); the total of 218 texts with 43,181 tokens of error annotation.

In each part of the corpus, we collected data on the use of specific tags labeling student errors, and separately—on the use of highest-level general tags used by annotators. As stated above, the tag to be assigned has to be as specific as possible, and a higher-level (more general) tag can be used in one of the two cases—when there is no further division (for example, there no "Singular" or "Plural" tags for nouns—we only have a more general "Noun number" tag), or when the use of one more general tag simplifies the use of three or more specific tags of the same level. The example of the latter case is the following:

(4) *The almost equal number of increasing international graduates was observed… >The almost equal increase in the percentage of international graduates was observed…* (text 6)

An annotator can either use three specific discourse tags to show the errors made—"Coherence", to change *number* for the word *percentage*; the same tag for the change from *increasing* to *increase* **(NOT a vocabulary error!)**, and "Absence of a component in a clause or sentence" to add preposition *in* to the combination *increase in the percentage*, or choose to use one general tag—"Discourse" to signify the overall change.

Figs. 3–5 below demonstrate the variation in annotation statistics in three areas of REALEC:

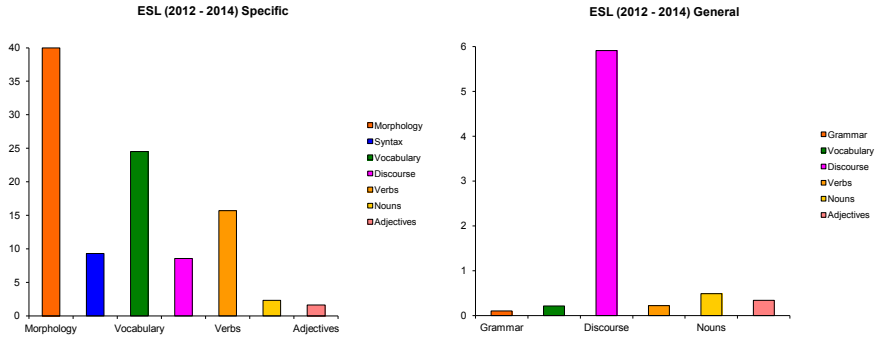


Fig. 3. Variation in the use of error tags in ESL (the initial learner corpus)

It can be concluded from the graph on the right that general tag DISCOURSE was applied to the overwhelming majority of cases when annotators could not classify errors as grammar or vocabulary, and also that there was insufficient subdivision of discourse errors. Correspondingly, we worked towards eliminating these deficiencies by adding more discourse tags and working out specific approaches to annotating discourse errors. As a result, in the more recent addition to the corpus the distribution of tags assigned by annotators is more even:

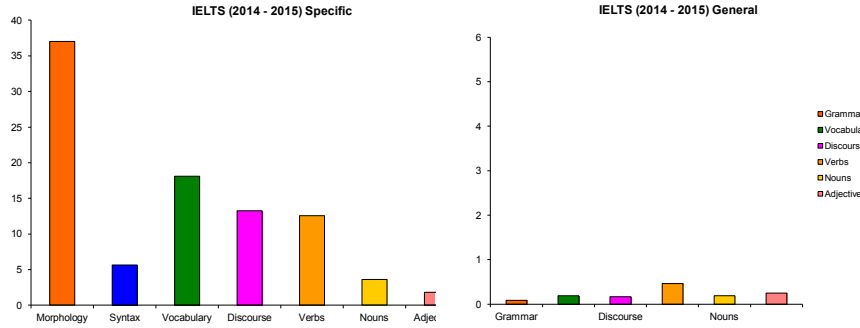


Fig. 4. Variation in the use of error tags in IELTS (the collection of examination essays in REALEC)

And finally, in the most recent texts added to REALEC—the essays written in 2015–2016 in preparation for their IELTS examination by the current students, who annotate themselves the errors that their instructors outline for them—there is only one case when a general tag was applied—it is the example very similar to the one discussed in (4) above:

(5) *It should be noted that the poorest group of poor people spends less on petrol—nearly 4 percent>It should be noted that the percentage of money spent on petrol by the poorest group of poor people in the two countries is very different.*

Instead of assigning three discourse tags—"Tautology" (because the same figure for the same group was given in the previous sentence), "Absence of the necessary information or detail" for the need to add in which country/countries, and "Coherence" for the need to talk about the difference for the two countries—the annotator decided to assign just one general tag—"Discourse," and for this single example of the use of high-level tag no graph on the right is presented in Fig. 5.

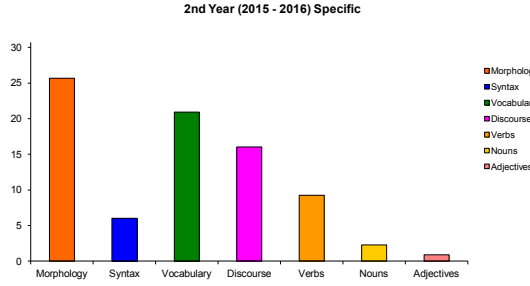


Fig. 5. Variation in the use of specific error tags in the current area of REALEC

To observe more inter-rater differences in REALEC annotation practices, we carried out the experiment recently (below referred to as Experiment 2), in which 12 annotators were given the task to annotate the same text about 350 words long. All

annotators were familiar with the annotation workflow, even if to a different degree, and the research interest was to list points of disagreement of different kinds.

The total number of error spans marked in this text was 156. Of them, 57 were spotted by no more than 2 annotators, 23 were spotted by only 3 annotators, 30 errors were marked by at least 10 annotators of the 12 participants, and they all chose the same tag for these spans, and 6 areas spotted by at least 10 annotators were marked with different tags. What is left is 40 tags noticed by 4 to 9 annotators, and there are 19 among them in which the annotators agreed in their choice of tags (for the convenience of reference called in the graph "Part agree"). Fig. 6 shows the distribution of the spread of annotation decisions across the 12 annotators in the experiment.

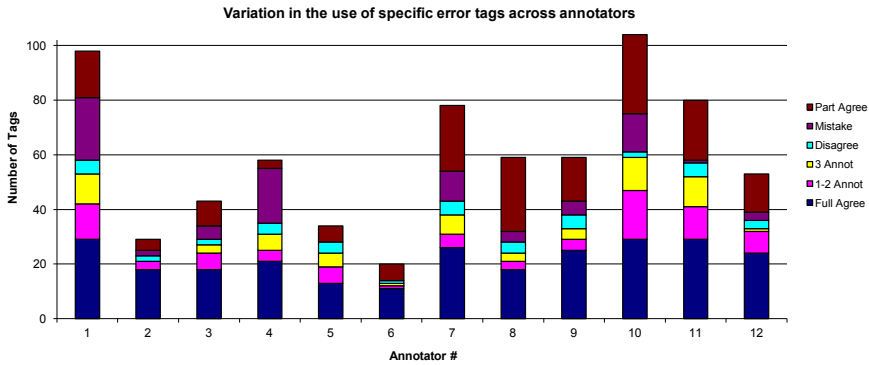


Fig. 6. Variation in the use of specific error tags by annotators in Experiment 2

To conclude, two annotation experiments demonstrated adequate reliability in the use of tags by REALEC annotators and in their approach to complicated errors. Hopefully, by increasing uniformity in annotation practice we will be able to approach automatisation of tagging in the learner corpus of student written works and, as a result, get closer to partially automated evaluation of student essays (as is indicated in McEnery & Xiao 2011). The corpus itself is a valuable pedagogical tool—for one, it provides a variety of possibilities for EFL instructors to create automated and semi-automated training exercises, as well as progress and placement tests on the basis of the mistakes annotated in learner texts in the corpus. The main feature of such exercises and tests is going to be their precision in targeting sharply at eliminating the specific mistakes that a particular group of learners is prone to making.

References

1. *Artstein R. & Poesio M.* (2008) Artstein, Ron and Massimo Poesio, Massimo Inter-coder Agreement for Computational Linguistics in Computational Linguistics 34(4), 2008, pp. 555–596.
2. *Bayerl P.* (2007) Bayerl, Petra Saskia Bayerl Identifying Sources of Disagreement: Generalizability Theory in Manual Annotation Studies in Computational Linguistics, 33(1), 2007, pp. 3–8.

3. *Braun S.* (2006) Braun, Sabine ELISA—a pedagogically enriched corpus for language learning Purposes in Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods Frankfurt/M: Lang, 2006, pp. 25–47
4. *Bryant Ch. & Hwee Tu Ng* (2015) Bryant, Christopher & Ng, Hwee Tou How Far are We from Fully Automatic High Quality Grammatical Error Correction?-in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 2015, pp. 697–707
5. *Chodorow, M. et al.* (2010) Chodorow, Martin, Gamon, Michael, Leacock, Claudia, & Tetreault, Joel Rethinking Grammatical Error Annotation and Evaluation with the Amazon Mechanical Turk—in Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 45–48
6. *Chodorow et al.* (2014) Chodorow, Martin, Gamon, Michael, Leacock, Claudia, & Tetreault, Joel. Automated Grammatical Error Detection for Language Learners 2014– in Series Title: Synthesis Lectures on Human Language Technologies Publisher: Morgan & Claypool Publishers, 2014
7. *Diez-Negrillo A. & Fernandez-Dominguez J.* (2006) Diez-Negrillo, Ana & Fernandez-Dominguez, Jesus Error-Tagging Systems for Learner Corpora—in Revista española de lingüística aplicada, # 19, pp. 83–102
8. *Glaznieks et al.* (2014) Glaznieks, Aivars, Nicolas, Lionel, Stemle, Egon, Abel, Andrea & Lyding Verena Establishing a Standardised Procedure for Building Learner Corpora in Journal of Applied Language Studies, vol. 8, 3, 2014, 5–20
9. *Granger S.* (2003) Granger, Sylviane 2003. The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research, TESOL Quarterly, 37, 3, p. 538–546
10. *Granger, Gilquin & Meunier* (2013) Granger, Sylviane, Gilquin, Gaëtanelle & Meunier, Fanny (eds) Twenty Years of Learner Corpus Research—Looking Back, Moving Ahead. Corpora and Language in Use—Proceedings 1, Louvain-la-Neuve, Presses Universitaires de Louvain, 2013
11. *Hovy E. & Lavid J.* (2010) Hovy, Eduard and Lavid, Julia Towards a "Science" of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics, in International Journal of Translation Studies 22(1), 2010: pp. 13–36.
12. *Hovy E.* (2015) Hovy, Eduard Corpus Annotation in Ruslan Mitkov (ed.) The Oxford Handbook of Computaional Linguistics Second Edition (2 ed.) Online publication November 2015
13. *Krippendorff K.* (2007) Krippendorff, Klaus Computing Krippendorff's Alpha Reliability available at http://web.asc.upenn.edu/usr/krippendorff/mwebreliability5.pdf
14. *Krippendorff K.* (2012) Krippendorff, Klaus. Content analysis: An introduction to its methodology. Sage, 2012.
15. *Kutuzov A., E. & O.* (2015) Kutuzov, Andrey, Kuzmenko, Elizaveta and Vinogradova, Olga Evaluating inter-rater reliability for hierarchical error annotation in learner corpora in the proceedings of 8th International Corpus Linguistics Conference, Lancaster 2015, pp. 211–214.

16. *Lee J. et al.* (2014) Lee, John, Yan Yeung, Chak, Zeldes, Amir, Reznicek Marc, Lüdeling Anke, and Webster, Jonathan CityU Corpus of Essay Drafts of English Language Learners: A Corpus of Textual Revision in Second Language Writing—electronic prepublication at https://corpling.uis.georgetown.edu/amir/pdf/annis_cityu_prepub.pdf
17. *McEnery, Tony and Richard Xiao* (2011) What corpora can offer in language teaching and learning. In Hinkel, E. (ed.), Handbook of Research in Second Language Teaching and Learning. London: Routledge.
18. *Poesio M., Bruneseaux F. & Romary L.*(1999) Poesio, Massimo, Bruneseaux, Florence & Romary, Laurent The MATE Meta-Scheme for Coreference in Dialogues in Multiple Languages In Proceedings of the ACL Workshop on Standards for Discourse Tagging, College Park, MD, pp. 65–74. Stroudsburg, PA: Association for Computational Linguistics.
19. *Pustejovsky J. & Stubbs A.* (2013) Pustejovsky, James & Stubbs, Amber (). Natural Language Annotation for Machine Learning. Sebastopol, CA: O'Reilly Media, 2013.
20. *Shtindlova B. et al.* (2014) Shtindlova, Barbara, Rosen, Alexandr, Hana, Jirka & Shkodova, Svatava—CzeSL—an error tagged corpus of Czech as a second language available at http://utkl.ff.cuni.cz/~rosen/public/2014-czesl-sgt-en.pdf
21. *Sorokin A. & Forsyth D.* (2008) Sorokin, Alexander, & Forsyth, David Utility Data Annotation with Amazon Mechanical Turk. In Proceedings of the First IEEE Workshop on Internet Vision at the Computer Vision and Pattern Recognition Conference (CPVR), 23–28 June Anchorage, AK, 1–8. Washington, DC: IEEE Computer Society, 2008.
22. *Stenetorp P. et al* (2012) Stenetorp, Pontus, Pyysalo, Sampo, Topić, Goran, Ohta, Tomoko, Ananiadou, and Tsujii, Jun-Ichi BRAT: A Web-Based Tool for NLP-Assisted Text Annotation. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 102–107. Stroudsburg, PA: Association for Computational Linguistics.
23. *Yannakoudakis H.* (2013) Yannakoudakis, Helen Automated assessment of English-learner writing. Cambridge, 2013 https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-842.pdf
24. *Wilcock* (2009) Wilcock, Graham. Introduction to Linguistic Annotation and Text Analytics. San Rafael, CA: Morgan and Claypool Publishers, 2009.

НОВЫЕ ИНТОНАЦИОННЫЕ КОНСТРУКЦИИ РУССКОГО ЯЗЫКА: РАЗРАБОТКА ТРАНСКРИПЦИИ[1]

Янко Т. Е. (tanya_yanko@list.ru)

Институт языкознания РАН, Москва, Россия

Задача работы — ввести в научный оборот две интонационные конструкции русского языка, не зафиксированные в русистике, и предложить для них нотацию. Первая конструкция обнаруживается в одном из региональных вариантов русского языка. Это подъем частоты на ударном слоге словоформы-носителя плюс продолжающийся подъем на заударных слогах. Вторая конструкция принадлежит русской разговорной речи. Рабочий массив записей фиксирует достаточно высокую частотность этой конструкции. Конструкция представляет собой падение в существенном диапазоне частот и в увеличенном временно́м промежутке плюс слабый подъем или вибрирующее изменение частот на заударных слогах или на второй части ударного слога при отсутствии заударных. Рассматриваются фонетические признаки конструкций, их фонологический статус и возможности использования системы просодической автосегментной транскрипции Дж. Пьерхамберт для их описания. Для анализа создан просодически размеченный рабочий массив записей устной речи. Работа иллюстрирована графиками изменения частоты основного тона, полученными с помощью системы анализа устной речи Praat.

Ключевые слова: интонация, интонационные конструкции, конструкции, не зафиксированные в описаниях, просодическая нотация, автосегментная транскрипция, русский язык, наддиалектный русский, одесский региональный вариант русского языка, звучащая речь, Praat

PROSODIC TRANSCRIPTION FOR THE NEW RUSSIAN PROSODIC CONSTRUCTIONS

Yanko T. E. (tanya_yanko@list.ru)

Institute of Linguistics, Moscow, Russia

This paper is aimed at investigating two Russian pitch accents never discussed in the Russian linguistics. These are: the gradual rise found in one the Russian regional variants, namely in Odessa Russian, and the prosody

[1] Исследование выполнено при поддержке Российского научного фонда (РНФ), проект №14-28-00130.

of breaking information into portions in standard literary spoken Russian. The gradual rise has a rising tone on the tonic syllable and gradually rising frequency changing on the post-tonic syllables. The prosody of breaking information into portions has a falling tonic syllable and slightly rising post-tonics. It also has a prolonged time of articulation. The tonal and temporal parameters of the pitch accents in consideration, their functions in discourse, and their phonological status are discussed. The criterion for the pitch accent to be viewed as an autonomous phonological unit of a language is whether the pitch accent has permanent means of expression and a stable function, or a limited set of functions in discourse. For describing the newly introduced pitch accents, the transcription based on the Pierrehumbert's autosegmental notation of prosody was used. For investigation, a minor working corpus of the Russian speech recordings was set up. It comprises two components. The first component of the corpus consists of short stories about Odessa told by the citizens, jokes, and funny stories. The second component includes recordings of friendly talks and radio conversational programs in standard Russian. The software program Praat was used in the process of analyzing the sounding data. The results presented here are exemplified by frequency tracings of records taken from the corpus.

Key-words: pitch accents, prosody, autosegmental transcription, Praat, frequency tracings, spoken language, standard Russian, Odessa regional variant of spoken Russian, newly introduced pitch accents

Данная работа преследует две цели. Одна из целей — ввести в научный оборот интонационную конструкцию русского литературного языка, не зафиксированную в русистике ранее. Обсуждение этой конструкции и ее документирование ставят перед нами вторую, более общую, цель — определения фонологического статуса интонационных конструкций, не зафиксированных в известных описаниях просодии тех языков, в которых они обнаруживаются, и разработки нотации для их документирования. Для иллюстрации этой проблемы будет рассмотрена еще одна конструкция, также не имевшая ранее точной атрибуции. Эта конструкция обнаруживается не в стандартном наддиалектном русском, а в одесском региональном варианте русского языка. С нее мы начнем обсуждение проблемы нотации для описания просодии. Таким образом, в разделе 1 ниже рассматривается не описанная ранее конструкция одесского регионального варианта русского языка, в разделе 2 — просодическая конструкция наддиалектного русского. Конструкцию одесского регионального варианта мы условно называем градуальным подъемом, потому что ее характеризует повышение частоты на ударном слоге словоформы-носителя конструкции плюс, при наличии заударных слогов, дальнейшее повышение частоты на первом заударном слоге, начиная от того уровня, который был достигнут в результате подъема на ударном слоге. На втором заударном слоге, если таковой имеется, частота также повышается, опять же начиная с того уровня, который был достигнут в результате подъема на первом заударном. Ср. изменение частоты основного тона, фиксирующееся на словоформе *продукция* в примере (1), рис. 1.

(1) *Естественно, не всегда распространялась полностью ПРОДУКЦИЯ…*

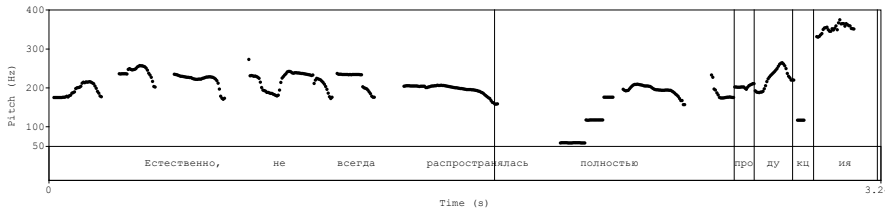


Рис. 1. Тонограмма примера (1)

Конструкцию из наддиалектного русского мы называем по ее функции: просодией эмфатического членения текста на отдельные кванты информации. Конструкцию иллюстрирует пример (2). Искомая просодическая конструкция фиксируется на словоформах *лет*, *тома*, *прецеденты* и *объяснения*. Конструкция характеризуется рельефным падением частоты основного тона на ударном слоге словоформы-носителя плюс подъем в небольшом диапазоне частот на заударных слогах, если они есть, или вибрация частоты на заударных, ср. движение тона на словоформах *прецеденты* и *объяснения* в примере (2). Если заударные слоги отсутствуют, финальное движение тона фиксируется на конечном фрагменте ударного слога, ср. в примере (2) движение тона на словоформах *лет* и *тома*.

(2) *Это много ЛЕТ, это исписаны ТОМА, это найдены ПРЕЦЕДЕНТЫ, это ОБЪЯСНЕНИЯ.*

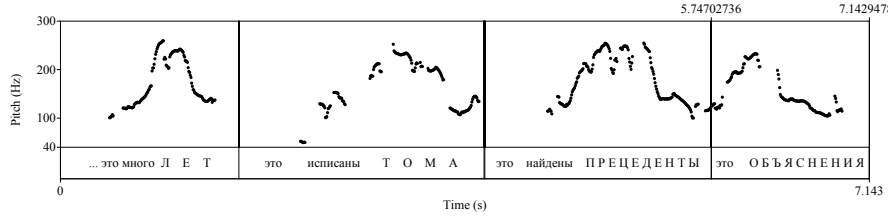


Рис. 2. Тонограмма примера (2)

При разработке нотации для этих конструкций встают следующие вопросы. Зафиксированы ли эти конструкции в просодии других языков и, если да, то какой фонологический статус они имеют в этих языках: служат ли они в этих языках отдельными просодическими единицами или играют роль вариантов для других просодий? Какие из существующих в интонологии просодических нотаций способны отразить отличительные признаки исследуемых акцентов?

По нашему мнению, градуальный подъем представлен в различных языках, и в качестве материала для сравнения с русскими региональными данными мы используем ниже немецкий пример и описание градуального подъема в соответствующей литературе.

В качестве основного критерия, определяющего фонологический статус конструкции, мы используем способность конструкции иметь особую дискурсивную функцию/ограниченный набор дискурсивных функций, отличных от функций других просодических единиц в системе.

Что же касается способов описания новых акцентов, вводимых в научный оборот, в связи с этим ниже потребуется обсуждение некоторых признаков, характеризующих интонационные конструкции. Так, при разработке нотации для нового акцента, наблюдаемого в наддиалектном русском, возникает необходимость фиксации особого параметра интонационных конструкций, состоящего в сохранении/потере вектора частоты, в общем случае приходящегося на заударные слоги, в применении к ситуации, когда заударные слоги отсутствуют. Это может быть усечение вектора частоты по принципу «отсутствуют заударные, отсутствует и соответствующее им движение тона» (как это происходит, например, в отсутствие заударных при реализации русского акцента типа ИК-3, по Е. А. Брызгуновой) или, наоборот, компрессия движения тона, при которой вектор движения на заударных не пропадает при их отсутствии, а сдвигается целиком «влево» — на вторую половину ударного слога (как это происходит при реализации русского акцента ИК-4). Проблема отражения в нотации способа наложения тональной кривой на сегментный материал с учетом ударности несущих слогов на примере различных акцентов русского и английского языков обсуждается в разделе 2 при разработке нотации для нового акцента наддиалектного русского языка.

Для анализа просодий был разработан малый исследовательский массив просодически размеченной звучащей речи. Массив состоит из двух основных компонентов. Первый компонент — рассказы одесситов о городе, одесские анекдоты, случаи из жизни, кулинарные рецепты. Второй компонент — это записи т. н. разговорных радиопередач и дружеских бесед, отражающих речь москвичей и жителей Санкт-Петербурга.

Кривые частоты основного тона получены с помощью системы анализа устной речи Praat.

1. Градуальный подъем

Анализ градуального подъема начнем с немецкого примера (3) из работы [Палько 2008]. В немецком языке градуальный подъем — это зафиксированный в описаниях акцент, он имеет определенные дискурсивные функции и, таким образом, его можно считать отдельной фонологической просодической единицей и использовать в качества образца для анализа.

(3) *Ich war in FRANKREICH. Dann war ich einen Monat in CHINA, mit einer FREUNDIN. Und dort haben wir ihre Eltern besucht, also, sie ist direkt aus CHINA.*

‘Я была во Франции. Потом я была один месяц в Китае, с одной подругой. А там мы посетили ее родителей, ведь она из Китая’.

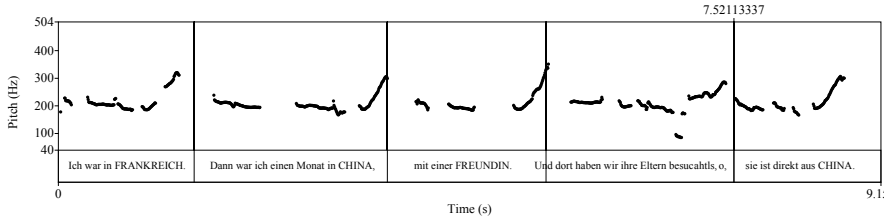


Рис. 3. Тонограмма примера (3)

В примере (3) мы наблюдаем подъем на ударном слоге словоформы *Frankreich* плюс дальнейший подъем на заударном слоге. Аналогичный контур фиксируется на словоформах *China*, *Freundin* и второй встречаемости словоформы *China*, носителях акцента незавершенности в клаузах, составляющих рассказ о поездках студентки во время каникул. Пример (3) иллюстрирует не простую незавершенность, а специализированную дискурсивную стратегию т. н. «рассказа по порядку» [Янко 2008: 203–206]. Рассказ по порядку предполагает, что повествование выстроено говорящим в определенной последовательности. В отличие от неупорядоченного повествования, которое строится из неконечных шагов, в рассказе по порядку не только текущий шаг, но и шаг, следующий за текущим (кроме последнего и предпоследнего), понимаются как неконечные. Иначе говоря, повествование выстроено в цепочку, состоящую — при последовательном применении стратегии рассказа по порядку — из более чем четырех шагов: первый шаг акцентируется как неконечный с указанием на то, что шаг, следующий за ним, будет тоже неконечный (в этом сущность просодии, маркирующей не пару, а цепь звеньев), аналогично акцентируются и следующие шаги повествования вплоть до предпоследнего шага. Предпоследний шаг имеет — в идеале — просодию предпоследнего звена в цепочке, а последний — просодию конца. В режиме рассказа по порядку, как правило, излагаются истории об экскурсиях с выстроенным в строгую последовательность порядком мероприятий (пример (3)), планы и расписания, презентации кулинарных рецептов и других технологий, предусматривающих выверенную последовательность шагов. Избрав стратегию рассказа по порядку, говорящий может говорить в соответствующем режиме не только о сущностях или событиях, упорядоченных во времени, в пространстве или в соответствии с внутренней логикой процесса, но и о любых незаконченных этапах повествования, просто делая вид, что этапы следуют один за другим: в таком случае как упорядоченный может представать не ход жизни, а только ход повествования. В русской интонационной системе функцию рассказа по порядку берет на себя акцент типа ИК-4, по Е. А. Брызгуновой [Русская грамматика 1982: 114] (падение на ударном слоге плюс подъем на заударных, если они есть; если заударные отсутствуют, восходящее движение тона фиксируется на конечном фрагменте ударного слога); о функциях ИК-4 дополнительно см. [Янко 2008: 209–218].

В одесском региональном варианте градуальный подъем, несмотря на достаточно частую встречаемость, в отличие от немецкого языка устойчивой дискурсивной функции не имеет. Обратимся к примеру (4) из «одесского» русского, так же, как и пример (3), повествующему о поездке в отпуск.

(4) *Мы покупаем билет до ХЕРСОНА, внутри доплачиваем ВОДИТЕЛЮ, нас везут до ГРАНИЦЫ, а там мы пешком переходим границу. Ну, у меня все спрашивают, потому что многие мои ДРУЗЬЯ — я очень многих людей знаю в ОДЕССЕ…*

Рис. 4. Тонограмма примера (4)

В примере (4) градуальный подъем наблюдается только на словоформе *Одессе*. Анализ этого и других примеров из одесского регионального варианта говорит о том, что градуальный подъем имеет здесь ту же функцию, что и восходяще-нисходящий акцент типа ИК-3, по Е. А. Брызгуновой, который широко представлен в одесском региональном варианте русского языка. Так, на словоформах *Херсона*, *водителю*, *границы* и *друзья* фиксируется акцент типа ИК-3 с подъемом на ударном слоге словоформы-носителя плюс падение до среднего уровня на заударных слогах, где они есть. Таким образом, в примере (4) и во многих других градуальный подъем попадает в один ряд с показателями незавершенности типа ИК-3. Это дает основания полагать, что словоформа *Одессе* в примере (4) несет на себе фонетический вариант ИК-3. В вопросе (5) из одесского русского также наблюдается градуальный подъем на словоформе *года*.

(5) *Самый короткий анекдот семьдесят второго ГОДА?*

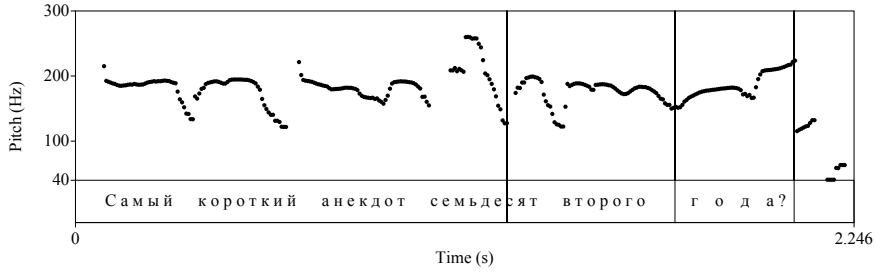


Рис. 5. Тонограмма примера (5)

Однако *да-нет*-вопросы с падением на заударных (т. е. типа ИК-3) в одесском региональном варианте также представлены, и гораздо более широко, чем вопросы с градуальным подъемом. Это опять же говорит о том, что в одесском

региональном варианте русского языка градуальный подъем — не отдельная просодическая единица, а свободный вариант акцента типа ИК-3.

Перейдем к проблеме нотации градуального подъема. Для немецкого градуального подъема в рамках автосегментной модели Дж. Пьерхамберт (см. [Pierrehumbert 1980]) предложена следующая транскрипция [Grice, Baumann, Benzmüller 2005]:

LH*H-^H%,

где буквы L и H обозначают низкий (L — low) и высокий (H — high) уровень частот основного тона, соответственно, пара LH обозначает восходящее движение тона, звездочка (*) — ударный слог, дефис (-) — безударный, в данном случае — заударный, слог. Знак процента (%) — конец конструкции, крышечка (^) — специальный знак, говорящий о том, что конец конструкции достигает частот, расположенных выше уровня, достигнутого в результате подъема на ударном слоге.

В работе [Gussenhoven 2005: 132] о голландском градуальном подъеме говорится, что, поскольку в голландском градуальный подъем так же, как, видимо, и в одесском, не имеет фонологического статуса, более высокие значения частот в финале заударных слогов, чем в начале, не имеет смысла отмечать в нотации: и начальный, и конечный уровень заударных помечаются просто как высокие: LHH%. См. также [Gussenhoven 2014]. Такое решение, во всяком случае для анализируемого одесского варианта русского языка, представляется нам неадекватным: независимо от фонологического статуса конструкции, при ее транскрибировании должно сохраняться ее отличие от акцента типа ИК-6, по Е. А. Брызгуновой, который имеет иные по сравнению с градуальным подъемом функции и тонально-темпоральную природу: у ИК-6 заударные сохраняются на высоком ровном или слабонисходящем (в силу естественного деклинационного спада) уровне, который практически не меняется от начала к концу заударных. Если бы в «одесском языке» акцент типа ИК-6 с высокими, но ровными заударными отсутствовал бы, картина поведения заударных внутри градуального подъема могла бы в транскрипции не детализироваться. Между тем акцент типа ИК-6 в одесском варианте есть, и, следовательно, нотация, не различающая тип заударных, для этого языка не подходит.

Что касается немецкого языка, в связи с которым была разработана нотация градуального подъема в виде LH*H-^H% (ср. также близкую по нотации модель L*H-^H%), то в немецком градуальный подъем явно имеет фонологический статус. Автосегментная модель немецкой просодии разработана авторами М. Грайс, С. Бауманном и Р. Бенцмюллером, представлена на сайте http://www.gtobi.uni-koeln.de/gm_tonale_repraesentation.html и в работах [Grice, Baumann, Benzmüller 2002], [Grice, Baumann 2005; 2007]. Авторы связывают функцию градуального подъема с формированием вопросов (*да-нет*-вопросов, переспросов и альтернативных вопросов). В соответствии же с нашими наблюдениями немецкий градуальный подъем, кроме вопроса, имеет также функцию рассказа по порядку, см. пример (3). Итак, поскольку у немецкого градуального

подъема есть устойчивые дискурсивные функции, градуальному подъему здесь следует приписать статус автономного элемента просодической системы.

В применении к градуальному подъему, который наблюдается в одесском региональном варианте русского языка (примеры (1), (4) и (5)), мы предлагаем также использовать нотацию LH*H-^H%, разработанную в немецкой интонологии, имея, между тем, в виду, что в одесском региональном варианте эта модель не представляет собой отдельной единицы и служит просодическим вариантом акцента типа ИК-3, формирующего контексты дискурсивной незавершенности и *да-нет*-вопросы.

2. Просодия эмфатического членения текста на отдельные кванты информации

Обратимся к конструкции литературного русского, проиллюстрированной выше примером (2). Эта конструкция позволяет говорящему достаточно независимо от синтаксической структуры членить предложения на фрагменты, каждый из которых соответствует отдельной позиции в аргументации говорящего. Применение конструкции характеризуется ее повтором, что формирует целый ряд последовательных компонентов текста. В примере (6) интересующая нас модель фиксируется на словоформах *семья*, *ребенок* и *люди*:

(6) *Молодая СЕМЬЯ, там… трехлетний РЕБЕНОК, да? э-э-э, любящие молодые ЛЮДИ…*

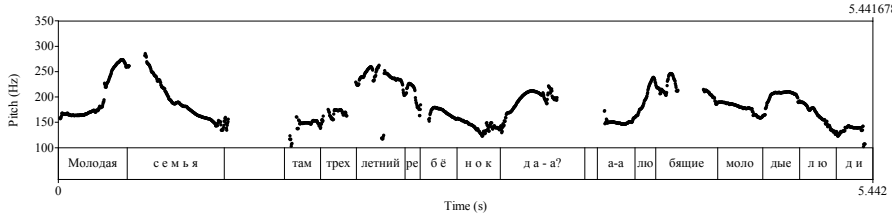


Рис. 6. Тонограмма примера (6)

На словоформе *семья*, не имеющей заударных слогов, наблюдается нисходящее движение тона в существенном диапазоне частот в начале ударного слога и вибрирующее движение в финале слога плюс растяжение артикуляции, которое накладывается на всю конфигурацию изменения частот. На словоформах *ребенок* и *люди* ударный слог несет на себе рельефное падение тона, а на заударный слог приходится небольшой подъем. Материал примера (6) и примера (2) выше позволяет заключить, что перед нами модель, которая подвергается компрессии в отсутствие заударных, она имеет вибрирующее или восходящее движение во второй фазе артикуляции, независимо от того, занимает вторая фаза заударные слоги или финальную часть конечного

ударного или единственного слога словоформы-акцентоносителя. Существенным признаком конструкции в целом служит удлинение артикуляции.

Интонационная модель, представленная примерами (2), (6) и примером (7) ниже, имеет известное фонетическое сходство с конструкциями, выделенными Е. А. Брызгуновой, а именно — с ИК-2, ИК-4 и ИК-5 ([Русская грамматика 1982: 98–115]), однако анализируемое здесь просодическое явление не сводимо в точности ни к одной из них. Нисходяще-восходящий аллофон рассматриваемой артикуляции — такой, как, например, на словоформах *ребенок* и *люди* в примере (6), имеет сходство с нисходяще-восходящим и также компрессируемым (в отсутствие заударных) акцентом ИК-4. Различие состоит в том, что анализируемый здесь акцент, во-первых, необязательно имеет отчетливо восходящую финальную часть, во-вторых, он имеет более крутое падение в начальной фазе (в случае с ИК-4 первая фаза может быть даже ровной низкой). В-третьих, рассматриваемая здесь конструкция имеет более длительное относительное время звучания (об ИК-4 см. [Русская грамматика 1982: 114–115]).

Имеется у анализируемой конструкции и общая черта с ИК-2. Это рельефное падение в начале ударного слога. Однако ИК-2 не имеет обязательной склонности к растяжению и финальному подъему частоты: в случае с ИК-2 заударный (если таковой имеется) финал конструкции произносится на низком уровне, а финал ударной части (если заударных нет) не отличается изменением частоты, как это происходит в рассматриваемой конструкции, имеющей по существу двухчастный характер. Таким образом, сходство с ИК-2 сводится к рельефному падению в центральной части артикуляции у обеих конструкций, между тем завершающая фаза у двух акцентов — различная.

Сравнительный анализ рассматриваемого здесь акцента и интонационной конструкции ИК-5 говорит о том, что заключительному падению в обоих случаях предшествует подъем тона, за которым следуют относительно ровные заударные вплоть до финального падения, однако между ними имеется отличие, которое состоит в следующем. В случае с ИК-5 конструкция имеет отчетливый двуцентровый характер, ср. ИК-5 в конструкциях с *какой* (*Какая гадость!* с подъемом на *какая* и падением на *гадость*). В обсуждаемой же здесь конструкции начальный подъем может рассматриваться как фонологически не значимый: он только обеспечивает достаточный уровень высоты, после достижения которого говорящий способен реализовать финальное падение в существенном диапазоне частот. Кроме того, конструкции ИК-5 не свойствен характерный для обсуждаемой конструкции восходяще-вибрирующий и растянутый заударный конец.

Таким образом, перед нами конструкция, в русистике, насколько нам известно, не зафиксированная. Этот факт может объясняться не очень высокой частотностью использования этой конструкции. В принципе, у нас достаточно материала, и из неподготовленной речи, и особенно из радиопередач, чтобы утверждать, что в речи говорящих по-русски данный акцент, безусловно, есть, однако не все носители используют этот акцент достаточно активно, а некоторые — не используют его вообще. Подсчет частотности данной конструкции не входил в нашу задачу, прежде всего потому, что имеющегося материала было достаточно для того, чтобы объявить о существовании незадокументированной интонационной

конструкции русского языка и сделать попытку ее анализа. Мы не исключаем, что наше внимание было привлечено распространением этой конструкции в т. н. разговорных передачах на радио, где говорящие стараются говорить настойчиво, акцентируя каждый аргумент особым акцентом повышенной эмфатичности и повторяя его в скандирующей манере. Анализ радиопередач говорит о том, что каждые час-полтора прослушивания дают в среднем одну-две серии употребления обсуждаемого акцента. В бытовой речи показатели частотности конструкции существенно ниже. В результате частотность рассматриваемой конструкции оказывается никак не меньше, чем общая частотность — при обращении к различным жанрам речи — частотности конструкции ИК-4 и неизмеримо выше, чем частотность ИК-7 (об ИК-7 см. [Русская грамматика 1982: 118–120]).

Дискурсивная функция анализируемого акцента состоит в подчеркнутом членении речи на фрагменты, соответствующие отдельным квантам информации: говорящий «дробит» речь, чтобы придать «мелким» квантам смысла автономное звучание. Так, в примере (7), отражающем неподготовленный фрагмент речи того же говорящего, от которого получен пример (2), значимая информация фактически отсутствует. Говорящий, выступающий в прямом эфире, не готов к ответу на вопрос радиослушателя, его речь звучит негладко, он смущен тем, что радиостанция, которую он представляет, потеряла в рейтинге, говорящий не знает, что ответить, он сбивается с мысли и по нескольку раз повторяет одни и те же слова. Однако стратегия дробления речи на кванты сохраняется: говорящий привык использовать эту стратегию и способен ее воспроизвести, используя минимум значимого сегментного материала.

(7) *Так вот, возвращаясь к рейтингам, тем не МЕНЕЕ, среди разговорных ПЕРЕДАЧ, в июле МЕСЯЦЕ, в июле месяце, среди разговорных ПЕРЕДАЧ, несмотря на то, что мы не восстановились, то есть мы к июлю прошлого восстановились, а вообще — не восстановились…*

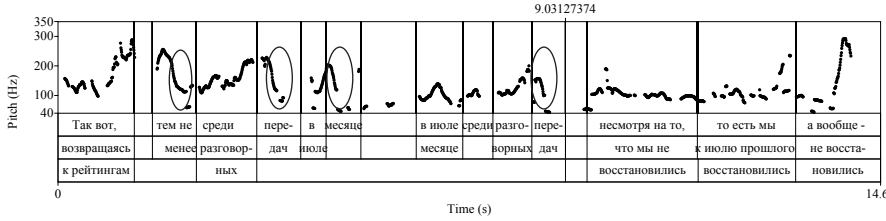


Рис. 7. Тонограмма примера (7)

Встает задача разработки нотации для рассматриваемой конструкции. Обратимся к нотации Дж. Пьерхамберт. В отличие от подхода Е. А. Брызгуновой, оперирующей цельными интонационными конфигурациями, которые имеют в русском языке фонологический статус, транскрипции американского диалекта и других языков, разработанные на основе нотации Пьерхамберт, имеют смешанный фонетико-фонологический статус и, в принципе, предусматривают презентацию

не конструкции в целом, а отдельных компонентов кривой изменения частоты тона от высоких частот (H) к низким (L) и наоборот. Большинство нотаций могут содержать в записи параметры как фонетического, так и фонологического характера. Так, например, восходящий акцент типа ИК-3, по Е. А. Брызгуновой, получает нотацию LH* или LH*-L в зависимости от того, имеются ли у словоформы-носителя акцента заударные слоги. В русском языке при наличии заударных частота на заударных, если они есть, при артикуляции ИК-3 резко падает; если заударных нет, артикулируется только восходящий компонент акцента LH*. Акценты, близкие по артикуляции к ИК-3, есть в большинстве языков, в частности, в английском, где, однако, снижение частоты на заударных отличается от русского меньшей крутизной падения. Возникает вопрос, как при данном типе записи отличить усекаемые акценты (ср. английский термин truncation), такие, как ИК-3, в которых движение тона на заударных при их отсутствии отбрасывается, от компрессируемых (ср. compression), таких, как ИК-4, в которых движение тона на заударных при их отсутствии смещается на финальную фазу ударного слога.

В описаниях английского языка для конструкции типа ИК-3 встречаются оба варианта записи LH* и LH*-L в зависимости от структуры сегментного материала [Cruttenden 1997: 60–61]. При этом запись LH* служит в качестве прототипического представителя акцента, а LH*-L — в качестве фонетической детализации ситуации, где имеются заударные слоги. Между тем, если принять в качестве фонологической записи вариант LH*, акцент типа ИК-3 нельзя будет отличить от другого вида подъема с ровными (или слабонисходящими) заударными типа ИК-6. Значит, ориентация на фонологическую запись с отсутствием заударных не позволяет различить как минимум два фонологически разных акцента. Итак, однозначного результата и — одновременно — фонологической ориентации нотации можно добиться только при указании на то, усечению подвергается акцент в отсутствие заударных или компрессии. Это делает запись более громоздкой, но и более последовательной с точки зрения описания.

Для отражения параметров рассматриваемой здесь конструкции мы предлагаем следующую нотацию: $\boxed{((HL^*\text{-}¡L\%):)\rightarrow\leftarrow}$. Фрагмент HL* символизирует падение тона на ударном слоге акцентоносителя, -L — низкий уровень заударного слога, перевернутый восклицательный знак ¡ говорит о том, что заударный слог, хоть и находится на относительно низком уровне, тем не менее, расположен выше той точки, которая была достигнута в результате падения на ударном слоге. В данной точке транскрипционной записи оказываются исчерпанными все известные нам знаки, которые используются в просодических автосегментных нотациях, и мы вводим два дополнительных знака — двоеточие, которое обозначает пролонгированную артикуляцию, и две направленные друг к другу стрелки, которые говорят о том, что перед нами конструкция, подвергающаяся компрессии, т. е. что низкий тон на заударных, которые, между тем, находятся выше точки падения на ударном, смещается на финал ударного слога при отсутствии заударных. Скобки используются для ограничения области действия параметра. Поскольку в данной записи имеется указание на вариативность конструкции в зависимости от наличия заударных слогов в слове-носителе, такую нотацию следует рассматривать как фонологическую.

***

В работе рассмотрены две интонационные конструкции русского языка, не зафиксированные ранее в русской интонологии. Это градуальный подъем, который представлен в одном из региональных вариантов русского языка, и просодия эмфатического членения информации на кванты в литературном наддиалектном русском. Документирование не засвидетельствованных ранее акцентов представляет самостоятельный интерес, кроме этого, возникает отдельная проблема описания конструкций и выяснения их фонологического статуса. В качестве критерия для признания фонологического статуса конструкции в языке использовалось наличие у конструкции устойчивой дискурсивной функции или ограниченного набора функций, который не совпадает с набором функций других конструкций языка. В соответствии с этим критерием было показано, что градуальный подъем в одесском региональном варианте не имеет фонологического статуса и служит вариантом интонационной конструкции типа ИК-3, по Е. А. Брызгуновой. Что же касается конструкции наддиалектного русского, было выяснено, что она имеет устойчивый набор тонально-темпоральных параметров и дискурсивную функцию эмфатического членения речи на кванты, соответствующие элементам аргументации говорящего. Соответственно, ее следует признать отдельной фонологической единицей русской просодии. Для описания рассмотренных конструкций была предложена нотация, основанная на автосегментной просодической транскрипции Дж. Пьерхамберт.

Литература

1. *Брызгунова Е. А.* (1982) Интонация, Русская грамматика, том 1, Наука, Москва, сс. 103–118.
2. *Палько М. Л.* (2008) Интонация незавершенности текста в немецком языке в сопоставлении с русским, Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог-2008». — М.: РГГУ, 2008. — Вып. 7 (14). — С. 416–419, available at: http://www.dialog-21.ru/digests/dialog2008/materials/html/65.htm.
3. *Янко Т. Е.* (2008) Интонационные стратегии русской речи в сопоставительном аспекте, Языки славянских культур, Москва.
4. *Cruttenden A.* (1997) Intonation, Cambridge University Press.
5. *Grice M., Baumann S.* (2002) Deutsche Intonation und GToBI, Linguistische Berichte, 191. 267–298.
6. *Grice M., Baumann S., Benzmüller R.* (2005). German Intonation in Autosegmental-Metrical Phonology, Prosodic Typology: The Phonology of Intonation and Phrasing, Oxford University Press, available at: http://www.coli.uni-saarland.de/publikationen/softcopies/Grice:19xx:GIA.pdf.
7. *Grice, M., S. Baumann* (2007) An Introduction to Intonation — Functions and Models, Non-Native Prosody. Phonetic Description and Teaching Practice, Trouvain J.,

Gut U. (eds.), Berlin, New York, De Gruyter, pp. 25–51, available at: http://www.gtobi.uni-koeln.de/lit/grice-baumann-int-func-models-revised-1107.pdf.

8. *Gussenhoven C.* (2005) Transcription of Dutch Intonation, Prosodic Typology: The Phonology of Intonation and Phrasing, Oxford University Press, available at: http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199249633.001.0001/acprof-9780199249633-chapter-5.
9. *Gussenhoven C.* (2014) Complex intonation near the tonal isogloss in the Netherlands Analysis, Prosodic Typology II: The Phonology of Intonation and Phrasing, Oxford: Oxford University Press.
10. *Pierrehumbert J.* (1980) The Phonology and Phonetics of English Intonation, MIT PhD Dissertation, available at: http://faculty.wcas.northwestern.edu/~jbp/publications/Pierrehumbert_PhD.pdf.

БАЗА ДАННЫХ МЕЖЪЯЗЫКОВЫХ ЭКВИВАЛЕНЦИЙ КАК ИНСТРУМЕНТ ЛИНГВИСТИЧЕСКОГО АНАЛИЗА[1]

Зализняк Анна А. (anna.zalizniak@gmail.com)

Институт языкознания РАН, Москва;
Институт проблем информатики ФИЦ ИУ РАН, Москва

Ключевые слова: база данных межъязыковых эквиваленций, параллельный корпус, русский язык, французский язык, лингвоспецифичные слова, дискурсивные слова, корпусная лингвистика, контрастивная лингвистика

A DATABASE OF CROSS-LINGUISTIC EQUIVALENCES AS AN INSTRUMENT OF LINGUISTIC ANALYSIS

Zalizniak Anna A. (anna.zalizniak@gmail.com)

Institut of linguistics RAS, Moscow;
Institute of Informatics Problems FRC CSC RAS, Moscow

The paper outlines the principles of creation of a Database of Russian language-specific units and their French equivalents and the possibilities of its use as a tool of linguistic analysis. The entry of the Database is a mono-equivalence (ME), i.e. a dyadic tuple, which consists of a Russian sentence including a language-specific unit and its French translation (automatically extracted from the Russian-French subcorpus of Russian National Corpus), including a functionally equivalent fragment (FEF) of the Russian language-specific unit. Both constituents of the ME are annotated with two-level characteristics, ensuring their faceted classification: “basic type” and “additional feature”. The paper indicates relevant quantitative parameters that can be extracted from such a database and can be accounted for in the analysis of language-specific units; it demonstrates that quantitative methods can be effectively used only in combination with proper methods of semantic analysis. The reliability of statistical data will increase with the extension of the volume of the parallel corpus.

Key words: database of cross-linguistic equivalences, parallel corpora, Russian language, French language, language-specific units, discourse markers, corpus linguistics, contrastive linguistics

---

[1] Статья написана при финансовой поддержке РФФИ, грант № 16-06-00339.

В данной статье излагаются принципы создания и возможности использования баз данных межъязыковых эквиваленций, формируемых на основе текстов параллельных корпусов, в качестве инструмента лингвистического анализа — на примере *базы данных лингвоспецифичных единиц русского языка*, созданной в ходе выполнения проекта «Контрастивное корпусное исследование специфических черт семантической системы русского языка», продолжением которого является проект исследования дискурсивных слов русского языка, которое проводится теми же методами и с использованием базы данных аналогичной структуры. Поскольку дискурсивные слова по преимуществу лингвоспецифичны, данный проект является естественным продолжением предыдущего. Основные принципы контрастивного корпусного исследования лингвоспецифичных единиц (ЛСЕ) были изложены в [Зализняк 2015]; в данной статье будут сделаны некоторые уточнения этих принципов, возникшие в ходе формирования базы данных, перечислены дополнительные (по сравнению с параллельными корпусами и другими типами электронных ресурсов) возможности, предоставляемые базой данных межъязыковых эквиваленций, а также продемонстрированы применяемые в наших проектах новые автоматизированные исследовательские технологии.

1. База данных лингвоспецифичных единиц русского языка и их французских функциональных эквивалентов

Как известно, непереводимость, или труднопереводимость — это важнейший отличительный признак лингвоспецифичных единиц. Однако до сих пор утверждения о «труднопереводимости» тех или иных единиц русского языка делались на основании сравнительного семантического анализа (см. в частности, [Зализняк, Левонтина, Шмелев 2005, 2012]), а также отдельных наблюдений над их употреблением, в том числе, частотностью (ср. о сравнительной частотности слов русск. *душа* и англ. *soul* в [Wierbicka 1992]) сопоставляемых единиц разных языков. С появлением корпусов параллельных текстов открылись новые перспективы изучения межъязыковой эквивалентности (ср. [Сичинава 2014, Соколова 2013, Алексеева 2007, Dobrovol'skij 2006, Добровольский 2013: 162–273, Добровольский 2015] и др. — по материалам английского, испанского и немецкого параллельных корпусов).

Однако эффективность контрастивного корпусного исследования существенно возрастает, если сделать еще один шаг, а именно, «надстроить» над параллельным корпусом базу данных, позволяющую фиксировать реально засвидетельствованные соответствия языковых единиц, снабженных характеристикой по множеству признаков. Концепция такой базы данных была создана силами нашего коллектива (см. [Loiseau et al. 2013, Бунтман и др. 2014, Зализняк и др. 2015]), и она была названа «надкорпусной базой данных» (НБД) [Кружков 2015]. Надкорпусная база данных межъязыковых эквиваленций формируется в автоматизированном режиме и предоставляет следующие новые возможности. Во-первых, в НБД для каждого употребления анализируемой

языковой единицы в тексте перевода уже выявлен ее «функционально-эквивалентный фрагмент» (ФЭФ)[2]; во-вторых, как анализируемой языковой единице, так и ее ФЭФ приписано значение признаков двухуровневой фасетной классификации, специально разрабатываемой отдельно для каждого типа языковых единиц. Кроме того, в НБД имеется возможность исправления ошибок выравнивания, которая в параллельных корпусах после загрузки текстов отсутствует.

Параллельный корпус, тексты которого используются при формировании надкорпусной базы данных лингвоспецифичных единиц (НБД ЛСЕ), состоит из двух сегментов: русские тексты с их переводом на французский язык (Р-Ф) и французские тексты с их переводом на русский язык (Ф-Р)[3]. Важно подчеркнуть, что в обоих сегментах нас интересуют единицы *русского языка* (список этих единиц составлен на основании Указателя лексем в книге [Зализняк, Левонтина, Шмелев 2012]). Соответственно, в Р-Ф сегменте мы ищем лингвоспецифичные единицы *в текстах оригинала*, а Ф-Р нас интересуют лингвоспецифичные единицы *в текстах перевода*. Перевод с русского языка на французский позволяет выявить те смысловые компоненты интересующей нас русской языковой единицы, которые определили выбор переводчиком того или иного переводного эквивалента. В случае «обратного» перевода в качестве свидетельства о семантике анализируемой единицы русского языка выступают признаки иноязычного текста, послужившие «стимулом» появления интересующего нас русского слова в переводе. При этом часто перевод *на* русский язык оказывается даже более информативным, чем перевод *с* русского языка — поскольку в этом случае лингвоспецифичное слово возникает в переводе как непосредственная реакция на смысловое задание, диктуемое иноязычным текстом (ср. [Шмелев 2015]).

Возможная лингвоспецифичность единиц французского языка нас в этом проекте не интересует: французский язык в обоих направлениях перевода служит не объектом, а лишь инструментом анализа, т. е. способом выявления скрытых семантических компонентов, содержащихся в значении анализируемых русских языковых единиц. В этом смысле наш метод контрастивного анализа был назван *унидирекциональным*. Это, однако, не означает, что создаваемая согласно этим принципам база данных не может служить различным переводческим и переводоведческим целям, а также целям сопоставительной лексикологии. Надкорпусная база данных лингвоспецифичных единиц русского языка является, тем самым, одновременно результатом и инструментом лингвистического анализа.

Входом НБД ЛСЕ является *моноэквиваленция* (термин был введен в [Loiseau et al. 2103]). Моноэквиваленция (в сегменте Р-Ф) — это двухместный кортеж (упорядоченная пара) вида: фрагмент текста на языке оригинала, содержащий

---

[2] Термин введен в [Добровольский и др. 2005]; более распространенный термин «модель перевода» (“translation pattern”), ср. [Hasselgård, Oksefjell (eds.) 1999, Сичинава 2014] имеет несколько более узкое значение.

[3] На момент 15.02.16 сегмент Р-Ф имел общий объем 2 104 900 словоупотреблений, 141 936 предложений; сегмент Ф-Р имел общий объем 470 458 словоупотреблений, 27 055 предложений. Работа по увеличению объема французского параллельного корпуса продолжается.

интересующую нас единицу русского языка — ее функционально-эквивалентный фрагмент (ФЭФ) в тексте французского перевода данного фрагмента. Моноэквиваленции автоматически объединяются в *полиэквиваленции* — в тех случаях, когда в БД имеется несколько переводов одного и того же исходного текста; ценность полиэквиваленции как инструмента анализа состоит в том, что она показывает варианты перевода языковой единицы *в одном и том же контексте*. Моноэквиваленции, построенные на основе Ф-Р сегмента корпуса имеют вид: «стимул перевода» (СП)[4], т. е. фрагмент французского текста, «реакцией» на которой служит появление в русском переводе интересующей нас языковой единицы — русский перевод. (Примеры моно- и полиэквиваленций см. ниже.)

В НБД ЛСЕ разработана двухуровневая система аннотирования языковых единиц, входящих в моноэквиваленции: «базовый вид» и «дополнительные признаки». Базовый вид — это классифицирующая категория; каждая анализируемая языковая единица принадлежит одному одноименному «базовому виду» (напр. слово *беда* — базовому виду *беда* и т. д.); базовые виды группируются в кластеры на основании принадлежности к определенной части речи.

Второй уровень — «дополнительные признаки» — обеспечивает фасетность классификации: дополнительные признаки различны по своей природе, и каждой анализируемой языковой единице приписываются значения нескольких признаков. Дополнительные признаки объединены в кластеры: в первом кластере помещены морфологические характеристики анализируемой языковой единицы, во втором — сведения о присутствующих в предложении зависимых элементах. В третьем кластере содержатся сведения о «конструкции» (в понимании Грамматики конструкций, т. е. сюда входит вся сочетаемость и идиоматика в широком смысле); например: *на* [*беду*]; [*беда*] *если*; [*беда*] *в том, что*; *жди* [*беды*] и т. п. В четвертом кластере помещаются характеристики типа предложения, в котором употреблена анализируемая языковая единица (вопросительное, восклицательное, побудительное, отрицательное; диалогическая реплика и др.).

Именно наличие этого второго уровня аннотирования обеспечивает новое качество информации, предоставляемой надкорпусной базой данных, по сравнению с параллельными корпусами, а также, например, такими переводческими ресурсами как Multitran или Linguee: НБД позволяет зафиксировать для исследуемой языковой единицы функционально эквивалентный фрагмент в тексте перевода (в том числе — грамматической природы, в том числе — нулевой) с учетом формально охарактеризованных признаков типа ее употребления.

Несколько иначе устроена система аннотирования единиц французского текста (т. е. ФЭФ). Главное отличие состоит в том, что список «базовых видов» — открытый, он составляется в процессе построения моноэквиваленций. Базовые виды ФЭФ сортируются по категориям: это основные части речи плюс следующие категории: Alia (сюда попадают слова всех остальных частей речи), Composita (многокомпонентные единицы), Sentential TS (ФЭФ, представляющие собой предложения, напр. *qu'est-ce que cela fait*; *on sait jamais* и т. п.), Grammatical

---

[4] Термин «стимул перевода» в том значении, в котором он применяется в наших исследованиях, введен в [Loiseau et al. 2013].

TS (грамматические средства передачи — например, значение русского глагола *успеть* может быть передано французской формой времени Passé antérieur).

Дополнительные признаки фасетной классификации единиц французского текста группируются в два кластера: Collocation и Grammatical. В кластер Collocation попадают словосочетания, содержащие некоторое слово из списка «базовых видов». Например, если в списке «базовых видов» имеется единица *coeur*, то словосочетание *au fond du coeur* попадет в кластер Collocation в виде *au fond du* [coeur]. В кластере Grammatical находятся грамматические характеристики ФЭФ.

Кроме того, некоторые признаки могут быть приписаны самой моноэквиваленции (изменение конструкции предложения по сравнению с оригиналом, необходимость учета контекста более широкого, чем данное предложение, существенное расхождение в способе лексикализации и др.)

Так, например, во фразе

> …во мне был заперт свет, который искал выхода, но только жег свою тюрьму, **не вырвался на волю** и угас (Гончаров. Обломов)

слово *воля* принадлежит базовому виду *воля* (кластер «существительное») и получает следующие значения дополнительных признаков: Sg (кластер 1); *вырваться на* [*волю*] (кластер 3), Neg (кластер 4).

Во французском переводе этой фразы находим ФЭФ для всего словосочетания (выделено жирным):

> …une lumière emprisonnée en moi cherchant une issue ne faisait que consumer sa prison et a fini par s éteindre **sans jamais recouvrer sa liberté.**

Базовый вид ФЭФ — *liberté*, **дополнительные признаки**: *recouvrer sa* [liberté] (**кластер** Collocation); *sans* INF (кластер Grammatical).

Соответствующая моноэквиваленция выглядит следующим образом (в первом столбце указан ее номер, во вором — шифр текста оригинала, в третьем — фрагмент текста, содержащий анализируемую ЛСЕ, в котором выделены сама ЛСЕ плюс ее релевантный контекст, в четвертом — базовый вид и доп. признаки данной ЛСЕ в данном предложении, в пятом — извлеченный из параллельного корпуса фрагмент перевода, в котором выявлен ФЭФ, в шестом — базовый вид и доп. признаки ФЭФ перевода):

604	ГОБ	во мне был заперт свет, который искал выхода, но только жег свою тюрьму, **не вырвался на волю** и угас.	**воля** < Neg > < Sg > < вырваться на [волю] >	une lumière emprisonnée en moi cherchant une issue ne faisait que consumer sa prison et a fini par s éteindre **sans jamais recouvrer sa liberté**	**liberté** < recouvrer sa [liberté] > < sans INF >

Другие примеры моноэквиваленций из сегмента Р-Ф:

892	ДПН	Нет, то досадно, что врут, да еще собственному **вранью** поклоняются.	**вранье** < Sg >	Non, ce qui est fâcheux, c'est qu'ils se trompent et qu'ils admirent pardessus le marché leurs propres **erreurs**.	**erreur** < Pl >
197	ДПН	**всё** гимнастикой **собираюсь лечиться**;	**собираться** < SubInf-IPF > < Pers1 > < V-IPF > < Pres > < всё >	**je me propose** toujours **de les soigner** par la gymnastique;	**se proposer** < SubInf >

Пример полиэквиваленции:

133	ГОБ	Обломову [...] хотелось бы, чтоб было чисто, [...] он [...] желал, **чтоб это сделалось как-нибудь так**, незаметно, само собой;	**как-нибудь** < так >	Oblomov eût [...] apprécié la propreté, mais à condition qu'elle s'installât d'elle-même, sans qu'il s'en aperçoive.	**ZERO**
171	ГОБ	Обломову [...] хотелось бы, чтоб было чисто, [...] он [...] желал, **чтоб это сделалось как-нибудь так**, незаметно, само собой;	**как-нибудь** < так >	Oblomov aurait [...] voulu que tout devînt propre [...] que la chose se fit insensiblement, et comme allant de soi.	**ZERO**

Примеры моноэквиваленций из сегмента Ф-Р:

2280	B99	pour donner une raison à sa présence à cette PPM (**et par extension** au sein de la société Madone)	**par extension**	чтобы оправдать свое присутствие на этом PPM (**а заодно и** в самой фирме «Манон»)	**заодно** < Adv >
2507	BIG	il lui arriva toujours, [...] de se mettre en fureur à cette observation,	**ZERO**	случалось все же, что он **вдруг** свирепел от этого замечания,	**вдруг** < Adv >
2614	MLH	et pleins de cloches qui sonnent **dans l'air bleu** des belles matinées	**air** < dans l'[air] bleu >	и там множество колоколов, которые звонят **в голубом просторе** прекрасного утреннего часа	**простор** < Sg > < в голубом [просторе]>

2. Корпусные методы анализа лингвоспецифичных единиц

В ходе работы с НБД ЛСЕ были выделены следующие пять основных типов отсутствия межъязыкового семантического изоморфизма; каждый тип обозначается реализующим его русским словом:

I. Асимметричное членение концептуальной области:
   a. подтип ЗНАТЬ (русск. *знать* vs. франц. *connaître — savoir*)
   b. подтип ПРАВДА-ИСТИНА (русск. *правда–истина* vs. франц. *verité*);
II. Тип САМОВАР (переводной эквивалент отсутствует; используется заимствованное слово или описательное определение);
III. Тип БАБУШКА (имеется один преимущественный вариант перевода, но он неточный);
IV. Тип РОДНОЙ (переводной эквивалент отсутствует; имеется несколько приблизительно равновероятных вариантов перевода, все неточные);
V. Тип РАЗЛУКА (имеется один преимущественный переводной эквивалент, имеющий более точное соответствие в русском языке): *разлука — séparation* (ср. *расставание*); *беда — malheur* (ср. *несчастье*); *грозный — menaçant* (ср. *угрожающий*); *тоска — angoisse* (*ср. тревога*).

С точки зрения возможностей применения количественных методов наибольший интерес представляют три последние категории, именно о них и будет идти дальше речь. Можно назвать следующий ряд характерных признаков, указывающих на вероятную лингвоспецифичность слова, которые могут быть установлены при помощи автоматизированных процедур, осуществляемых в НБД, «надстроенной» над параллельным корпусом. Перечислим эти признаки (их список и формулировки существенно уточнены по сравнению с теми, которые были приведены в [Зализняк 2015]).

А именно, на возможную лингвоспецифичность слова указывает:

- при переводе с русского языка:

(1) наличие большого количества ФЭФ, в том числе — имеющих приблизительно равную частотность (ср. ниже);

(2) наличие многокомпонентных ФЭФ (ср.: *обидно — en avoir gros sur le coeur; родной — membre de la famille*), а также ФЭФ, состоящих из двух квазисинонимов; ср. для слова *грозный*:

> *И он, как* ***грозный*** *учитель, глядел на прячущегося ребенка —*
> *Il le regardait comme un maître* ***sévère*** *regarde,* ***menaçant****,*
> *un enfant qui se cache.* (Гончаров. Обломов);

для слова *обидно*:

> должно быть, ей очень ***обидно***… — Sans doute se sentait-elle très ***malheureuse***, très ***déçue*** (Гончаров. Обломов);

(3) слово остается без перевода (модель перевода — ZERO); ср. для слова *душа*:

Впрочем, он был ***в душе*** *добрый человек* (Гоголь. Шинель) — *C'était pourtant un brave homme*;

- при переводе на русский язык:

(4) большое количество «стимулов перевода» (СП), в том числе, имеющих приблизительно равную частотность; напр. для слова *беда* из 17 примеров в 15 имеются разные СП;

(5) наличие многокомпонентных СП; напр. для появления слова *душа* в русском переводе имеются следующие СП во французском оригинале: *au fond de l'âme, au fond du coeur, dans la serenité, par un geste naturel, autant qu'ils veulent* и еще 10 различных многокомпонентных СП;

(6) отсутствие какой-либо единицы, которая «стимулирует» появление анализируемой русской ЛСЕ (стимул перевода — ZERO), ср. для слова *родной*:

Paris redevenait […] *ma ville* (P. Modiano. Quartier perdu). *Париж вновь становился* […] ***моим родным городом***.

Посмотрим теперь, как ведет себя слово *тоска* относительно перечисленных выше признаков лингвоспецифичности — на основании данных, полученных из БД ЛСЕ. Напомним, что *тоска* — одно из самых известных лингвоспецифичных русских слов, см. в частности [Wierzbicka 1992, Шмелев 2002][5].

Статистическая таблица, генерируемая базой данных (сегмент Р-Ф), выглядит следующим образом (цифра напротив слова указывает на количество примеров, в первом столбце указан французский переводной эквивалент, во втором — сколько раз он встретился, в третьем — какой процент от общего количества ФЭФ для данного русского слова это составляет):

тоска	111

ФЭФ французского языка	**Кол-во МЭ**	**% в группе**
angoisse	45	40,18%
détresse	19	16,96%
tristesse	8	7,14%
ennui	5	4,46%
angoissé	3	2,68%

[5] В работе [Соколова 2013] обнаружено 24 модели перевода на испанский язык слова тоска в рассказах А. П. Чехова; в работе [Сичинава 2014] обнаружено 22 модели перевода этого слова на английский в англо-русском подкорпусе НКРЯ.

ФЭФ французского языка	Кол-во МЭ	% в группе
nostalgie	3	2,68%
chagrin	3	2,68%
ZERO	2	1,79%
mélancolie	2	1,79%
désespoir	2	1,79%
douleur	1	0,89%
tourment	1	0,89%
triste	1	0,89%
mélancolique	1	0,89%
accablement	1	0,89%
s'ennuyer	1	0,89%
hypocondrie	1	0,89%
stupeur	1	0,89%
inquétude	1	0,89%
idées noires	1	0,89%
déprimé	1	0,89%
déprime	1	0,89%
se sentir triste	1	0,89%
angoisser	1	0,89%
ennui, tristesse	1	0,89%
tristesse et nostalgie	1	0,89%
ennui teinté d'affliction	1	0,89%
ennuyer	1	0,89%
désolé	1	0,89%

Признак (1) — количество разных ФЭФ — 27: соотношение наиболее частотных моделей перевода: 40%, 17%, 7%, 4,5%; очевидно преобладание одного — *angoisse*.

Признак (2) — наличие многокомпонентных ФЭФ:

> *…не столько от боли,* [...] *сколько от* ***тоски*** *— pas tellement sous l'effet de la douleur physique* [...] *que* ***d'ennui, de tristesse****..;* (Л. Толстой. Смерть Ивана Ильича);
>
> *всё* ***тоска — tristesse et nostalgie*** *sans fin* (Л. Толстой. Смерть Ивана Ильича,);
>
> *Тупая* ***тоска*** *— Cette* ***sorte de stupeur*** *obtuse et* ***chagrine*** (Л. Толстой. Смерть Ивана Ильича);
>
> ***Тоска*** *проглянула в лице Лужина — Une expression* ***d'ennui teinté d'affliction*** *passa sur le visage de Loujine.* (Достоевский. Преступление и наказание).

Признак (3): Модель перевода — ZERO представлена лишь в двух случаях.

В Ф-Р сегменте слово *тоска* встречается всего 11 раз, там преобладание варианта *angoisse* еще более существенное, но общее количество примеров слишком мало для каких-либо статистических выводов.

Следует упомянуть еще один существенный количественный параметр, который может быть извлечен из НБД ЛСЕ: он касается преимущественного переводного французского эквивалента; соответственно:

а) в сегменте Р-Ф: из каких русских «стимулов перевода» возникает выявленный преимущественный эквивалент во французском языке;

б) в сегменте Ф-Р: как переводится этот преимущественный эквивалент на русский язык.

Про преимущественный эквивалент для русского *тоска* — франц. *angoisse* — пока можно сказать, что в Р-Ф сегменте корпуса оно встречается 89 раз (из них 45, т. е. половина, соответствуют в русском слову *тоска*). В Ф-Р сегменте корпуса *angoisse* встречается 17 раз, из них оно переведено на русский язык словом *тоска* 8 раз, кроме того имеются варианты *тоскующий*, *тоскливо*, *тоскливая тревога*.

Из всего этого, как представляется, можно сделать вывод, что русское слово *тоска* и франц. *angoisse* имеют довольно существенную область совпадения. Другими словами, лингвоспецифичность слова *тоска* в паре «русский — французский» достаточно низкая. Эти данные, конечно же, должны быть верифицированы на большем материале.

Лингвоспецифичность существительного *обида*, а также глаголов *обидеться*, *обижаться* и в особенности предикатива *обидно* много обсуждалась в литературе (см. [Зализняк 2000, Dobrovolskij 2006, Протасова 2006, Апресян 2011]. В [Апресян 2011] приводятся примеры из Набоковской «Лолиты» в английском оригинале и авторском переводе на русский, где русское слово *обида* используется как переводной эквивалент пяти несинонимичных английских слов, а также примеры перевода глаголов *обидеться, обижаться* с русского на английский, выполненный билинвами (8 типичных контекстов употребления; все переводятся по-разному).

В НБД ЛСЕ в сегменте Р-Ф слово *обида* встретилось 56 раз. Представлено 27 вариантов перевода, из них 19 встречаются по одному разу; в 9% случаев оставлено без перевода; наиболее частотный перевод (*offense*) составляет 23%, т. е. меньше четверти случаев.

Если взять нелингвоспецифичное слово *любовь*, то картина будет разительно отличаться, а именно, в 85% случаев употребления данного слова в переводе ему соответствует одно и то же слово — *amour*.

Дискурсивное слово *авось* в 43% случаев переведено как *peut-être* (т. е. с утратой некоторой части смысла); следующий по частотности вариант — ZERO (23%); далее следуют варианты перевода, в которых отражено представление о желательности и малой вероятности обсуждаемого события (*espérer* — 10%, *avec un peu de chance*, букв.: ‘если повезет’ — 10%) и несколько единичных переводов, по-разному выражающих идею надежды на нечто маловероятное и непредсказуемость будущего: *peut-être bien, au petit bonheur, on sait jamais*.

Можно сказать, что в целом множество вариантов перевода достаточно полно отражает набор семантических компонентов этого слова, лингвоспецифичность которого возникает за счет *комбинации* нескольких компонентов: непредсказуемости хода вещей, легкомысленного расчета на осуществление желаемого положительного события и, одновременно, равнодушия к возможному провалу (о слове *авось* см. в частности [Wierzbicka 1992: 433–435]).

Для дискурсивного слова *как-нибудь* показателен, прежде всего, факт преобладания варианта ZERO (37%), а также общее количество разных вариантов перевода, в том числе, встретившихся по одному разу. Это множество вариантов перевода рисует богатую картину смысловых компонентов и возможных вариантов значения русского слова *как-нибудь* — одновременно подтверждая его лингвоспецифичность, определяемую, прежде всего уникальной комбинацией этих компонентов в одном типе употребления.

3. Заключение

Из сказанного могут быть сделаны следующие выводы.

1. Надкорпусные базы данных представляют собой потенциально весьма эффективный инструмент, который может быть использован:
   - для лингвистического анализа тех или иных единиц одного языка;
   - для проведения контрастивных исследований;
   - для совершенствования систем перевода — как, автоматического, так и авторского;
2. Количественные методы установления меры лингвоспецифичности существуют, но эффективно они могут применяться лишь в комбинации с методами собственно семантического анализа.
3. Для того чтобы результаты такого рода исследований были значимыми, нужны корпуса значительно большего объема (т. е. объем параллельного корпуса должен быть увеличен на один-два порядка, при этом он должен быть сбалансированным в отношении времени создания и жанра входящих в него текстов);
4. За те 150–200 лет, которые отделяют нас от эпохи «русской классической литературы» в русском языке произошли очень существенные семантические сдвиги, в особенности для тех языковых единиц, которые сегодня являются лингвоспецифичными: в подавляющем большинстве случаев в XIX веке их значение было иным, и оно не было лингвоспецифичным; так, почти все обсуждавшиеся выше слова претерпели существенную эволюцию в этом направлении: *тоска* (ср. *тоска за жизнь* в «Обломове»), *разлука, совестно, авось* и др. — ср. следующие два примера употребления выражения *без обиды*:

> …да так, что ни с того ни с сего сгреб кирпич и кинул в начальника, *безо всякой обиды* с его стороны (Достоевский. Преступление и наказание).

Жил папа *без обиды*, он считал, что это время было такое.
(С. Алексиевич. Время сэконд-хенд).

Факт увеличения лингвоспецифичности в ходе семантической эволюции уже отмечался исследователями (см., напр., [Шмелев 2001]), однако использование базы данных, надстроенной над параллельным корпусом, может здесь дать результаты более высокого уровня надежности. Таким образом, при должном расширении состава текстов параллельного подкорпуса НКРЯ (в котором на настоящий момент преобладают тексты XIX в.) диахронический аспект изучения лингвоспецифичных единиц также получит дополнительную экспериментальную основу.

Автор благодарит анонимных рецензентов, чьи ценные замечания были по возможности учены в окончательной версии статьи.

Литература

1. *Алексеева М. Л.* (2007), Русские реалии в разновременных немецких переводах романов Ф. М. Достоевского. Словарь-справочник. Екатеринбург.
2. *Апресян В. Ю.* (2011), Опыт кластерного анализа: русские и английские эмоциональные концепты. Часть 1 // ВЯ, №1, 2011.
3. *Бунтман Н. В., Зализняк Анна А., Зацман И. М., Кружков М. Г., Лощилова Е. Ю., Сичинава Д. В.* (2014), Информационные технологии корпусных исследований: принципы построения кросс-лингвистических баз данных // Информатика и ее применения. Т. 8, вып. 2. С. 98–110.
4. *Добровольский Д. О.* (2013), Беседы о немецком слове. М.: Языки славянской культуры, 2013.
5. *Добровольский Д. О.* (2015), Корпус параллельных текстов и сопоставительная лексикология // Труды института русского языка им. В. В. Виноградова. Вып. 6. М., 2015. С. 411–446
6. *Добровольский Д. О., Кретов А. А., Шаров С. А.* (2005), Корпус параллельных текстов: архитектура и возможности использования. // Национальный копус русского языка: 2003–2005. Результаты и перспективы. М., 2005. С. 263–296.
7. *Зализняк Анна А.* (2000), О семантике щепетильности (*обидно*, *совестно* и *неудобно* на фоне русской языковой картины мира) // Логический анализ языка. Языки этики. М., 2000. С. 101–118.
8. *Зализняк Анна А.* (2015), Лингвоспецифичные единицы русского языка в свете контрастивного корпусного анализа // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог'2015. М., 2015. С. 651–662.
9. *Зализняк Анна А., Левонтина И. Б., Шмелев А. Д.* (2005), Ключевые идеи русской языковой картины мира. Москва: Языки славянской культуры. М.
10. *Зализняк Анна А., Левонтина И. Б., Шмелев А. Д.* (2012), Константы и переменные русской языковой картины мира. Москва: Языки славянской культуры. М.

11. *Зализняк Анна А., Зацман И. М., Инькова О. Ю., Кружков М. Г.* (2015), Надкорпусные базы данных как лингвистический ресурс // Труды международной конференции» Корпусная лингвистика-2015». 22–26 июня 2015, Санкт-Петербург. СПб, 2015. С. 211–218.
12. *Кружков М. Г.* (2015), Информационные ресурсы контрастивных лингвистических исследований: электронные корпуса текстов // Системы и средства информатики. Т. 25. № 2. С. 140–159.
13. *Протасова Е.* (2006), Какого вкуса *обида*? // Integrum: точные методы и гуманитарные науки. Г. Нечипорец-Такигава (ред.) М.: «Летний сад», 2006. С. 270–288.
14. *Сичинава Д. В.* (2014), Использование параллельного корпуса для количественного изучения лингвоспецифичной лексики // Язык, литература, культура: Актуальные проблемы изучения и преподавания. Вып. 10. М., МАКС ПРЕСС, с. 37–44
15. *Соколова Л. В.* (2013), Способы передачи безэквивалентной лексики в переводах А. Чехова на испанский язык (на материале концепта «тоска») // Rafael Guzmán Tirado, Irina A. Votyakova (ed.). Tipología léxica. Granada, 2013. С. 191–196.
16. *Шмелев А. Д.* (2001), Некоторые тенденции семантического развития русских дискурсивных слов. // Русский язык: пересекая границы. Дубна, 2001. С. 266–279.
17. *Шмелев А. Д.* (2002), Русская языковая модель мира. Опыт словаря. М.: Языки славянской культуры.
18. *Шмелев А. Д.* (2015), Русские лингвоспецифичные лексические единицы в параллельных корпусах: возможности исследования и «подводные камни» // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог'2015. М., 2015.
19. *Dobrovol'skij D.* (2006), Zur kontrastiven Analyse kulturspezifischer Konzepte // Wörter-Verbindungen. Festschrift Jarmo Korhonen zum 60. Geburtstag. Hrsg. von Ulrich Breuer und Irma Hyvärinen. Frankfurt am Main etc.: Peter Lang, 2006. S. 31–45.
20. *Dobrovol'skij D., Pöppel L.* (2015), Corpus perspectives on Russian discursive units: semantics, pragmatics, and contrastive analysis // Yearbook of Corpus Linguistics and Pragmatics 2015. New York etc.: Springer, 2015, pp. 223–241.
21. *Hasselgård, H., Oksefjell, S.* (eds.) (1999), Out of Corpora: Studies in Honour of Stig Johansson. Amsterdam — Atlanta, GA: Rodopi.
22. *Kruzhkov M., Buntman N. V., Loshchilova E. J. Sitchinava D. V., Zalizniak Anna A., Zatsman I. M.* (2014), The database of Russian verbal forms and their French translation equivalents // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2014. С. 275–287.
23. *Loiseau S., Sitchinava D. V., Zalizniak Anna A., Zatsman I. M.* (2013), Information technologies for creating the database of equivalent verbal forms in the Russian-French multivariant parallel corpus // Информатика и ее применения, 2013. Том 7, вып. 2. С. 100–109.
24. *Wierzbicka A.* (1992), Semantics, Culture, and Cognition. Universal Human Concepts in Culture-Specific Configurations. N. Y.; Oxford: Oxford Univ. Press.

К ВОПРОСУ ОБ АСПЕКТУАЛЬНОМ СТАТУСЕ КОНАТИВНЫХ ПАР В РУССКОМ ЯЗЫКЕ: ПОЧЕМУ *ИСКАТЬ* НЕ МОЖЕТ ОЗНАЧАТЬ *НАЙТИ*?[1]

Зализняк Анна А. (anna.zalizniak@gmail.com)
Институт языкознания РАН, ФИЦ ИУ РАН, Москва

Микаэлян И. Л. (irina-mikaelian@yandex.ru)
The Pennsylvania State University, USA

Ключевые слова: русский язык, аспектология, видовая пара, конативные глаголы, критерий Маслова

ON THE ASPECTUAL STATUS OF CONATIVE PAIRS IN RUSSIAN: WHY *SEARCH* CANNOT MEAN *FIND*?

Zalizniak Anna A. (anna.zalizniak@gmail.com)
Institute of Linguistics, RAS, Moscow, Russia

Mikaelian I. L. (irina-mikaelian@yandex.ru)
The Pennsylvania State University, USA

The discussion of the so-called conative pairs in the Russian language, i. e. pairs including an imperfective and a perfective verb where the former expresses the idea of an attempt to reach a goal while the latter designates a successful fulfillment of that goal, has a long tradition.
However, the aspectual status of such pairs remains unclear. In particular, the aspectual status of the verbs *iskat'*$_{ipf}$ — *najti*$_{pf}$ (*to search — to find*) has not found consensus among scholars of the Russian language. The paper provides answers to the following questions: Why the verbs *iskat' — najti* do not function as an aspectual pair in Russian? Why however Russian speakers perceive this pair as an aspectual one? How the non-aspectual pair *iskat' — najti* is related to conative aspectual pairs *lovit' — pojmat'* и *reshat' — reshit'?* How the non-aspectual pair *iskat' — najti* is related to telic aspectual pairs *razyskivat' — razyskat'* and *otyskivat' — otyskat'*?

Key words: Russian language, aspectology, aspectual pair, conative verb, Maslov criterion

[1] Статья написана при поддержке РГНФ, грант №15-04-00507.

Обсуждение аспектуального статуса конативных пар (т. е. пар глаголов, где глагол несов. вида включает смысловой компонент ‘попытка [достижения результата X]’, а глагол сов. вида — компонент ‘успех [в достижении этого результата]’), имеет достаточно давнюю традицию — см. [Маслов 2004, Апресян 1980, Гловинская 1982, 2001, Булыгина, Шмелев 1997, Зализняк, Шмелев 2000, Плунгян 2001, Падучева 2004] и др. Как справедливо отметил Ю. С. Маслов в известной статье 1948 г., «[…]противоположение ‘попытка’ — ‘успех’, одно из важных противоположений в системе функций видовых пар, может выражаться в русском языке и независимо от вида, посредством других соотносительных глагольных пар, не видовых, а чисто лексических» [Маслов 2004: 88]. Однако вопрос о том, какие из глагольных пар с данным семантическим соотношением являются видовыми, а какие — нет, до сих пор далек от своего разрешения.

Особое место в этой проблематике занимает пара глаголов *искать — найти*. Более того, можно сказать, что пара *искать — найти* представляет собой некий казус в русской аспектологической теории и преподавательской практике, а также некую семантическую загадку, выходящую за рамки русского языка. Действительно, еще в работе [Маслов 2004: 85] было убедительно показано, что данная пара глаголов не может считаться видовой, поскольку глагол *искать* ни в каких контекстах не может принимать то же значение, что глагол *найти* — а общность значения глаголов НСВ и СВ признается основополагающим требованием для установления видовой пары всеми без исключения аспектологами. Между тем различные грамматические описания русского языка, от школьных пособий до теоретических работ, регулярно приводят пару *искать — найти* в списке супплетивных видовых пар, наряду с образцовыми супплетивными парами *ловить — поймать*, *брать –взять*, *класть — положить* и *говорить — сказать*.

Ср. следующий пример (из учебника русского языка для 6-го класса):

> «Глаголы совершенного и несовершенного вида часто образовывают видовые пары. Видовые пары — это глаголы разных видов, которые имеют одинаковое лексическое значение. Например: *вычеркнуть* (совершенный вид) — *вычеркивать* (несовершенный вид); *достигнуть* (совершенный вид) — *достигать* (несовершенный вид); *удвоить* (совершенный вид) — *удваивать* (несовершенный вид). Большинство глаголов, которые образуют пары, имеют одинаковый корень. Исключениями являются такие видовые пары как: *взять* (совершенный вид) — *брать* (несовершенный вид); *найти* (совершенный вид) — *искать* (несовершенный вид); *поймать* (совершенный вид) — *ловить* (несовершенный вид)». http://www.nado5.ru/e-book/vidy-glagola[2]

---

[2] Аналогичные утверждения находим также, например, в [Шмелева 2011: 7, Озеров 2006: 27]. Ср., впрочем, следующий тезис: «Исключительно в учебных, чисто прагматических целях иногда в одну видовую пару объединяются глаголы, которые, строго говоря, парой не являются. Так, наряду с парой *находить — найти* дается пара *искать — найти* для того, чтобы учащийся имел возможность увидеть различие в значении совершенного и несовершенного вида.» [Шустикова (ред.) 2012: 5].

Если учесть, что пара *искать — найти* рассматривается как видовая также в книге Ю. Д. Апресяна «Лексическая семантика» [Апресян 1974/1995: 83–84], становится очевидно, что такое представление должно иметь под собой некое когнитивное основание.

Задача данной работы в том, чтобы попытаться ответить на следующие вопросы:

1) Почему интуиция носителя языка воспринимает пару *искать — найти* как видовую?
2) Почему пара *искать — найти* все же не является видовой — или, другими словами, почему глагол *искать* не может обозначать событие ‘найти’? Этот вопрос включает следующие два:
    a) Как не-видовая пара глаголов *искать — найти* соотносится с видовыми парами *ловить — поймать* и *решать — решить*; а именно, в силу каких семантических причин глагол *искать* не обладает способностью обозначать событие ‘найти’ — в отличие от глаголов *ловить* и *решать*, которые могут принимать значение, соответственно, ‘поймать’ и ‘решить’?
    b) Как не-видовая пара глаголов *искать — найти* соотносится с видовыми парами *разыскивать — разыскать* и *отыскивать — отыскать*? Есть ли здесь какое-то семантическое различие, объясняющее способность глаголов *разыскивать* и *отыскивать* обозначать событие, соответственно, ‘разыскать’ и ‘отыскать’ — в отличие от глагола *искать*, неспособного брать на себя значение ‘найти’?[3]

1. Ответ на вопрос, почему *искать — найти* хочется считать видовой парой, в первом приближении состоит в том, что для говорящего на русском языке глаголы *искать* и *найти* прочно связаны — прежде всего, результативным отношением, аналогичным отношению между глаголами *решать* и *решить*, *ловить* и *поймать*[4]. Эта связь подкрепляется как минимум тремя следующими обстоятельствами.

Во-первых, тем фактом, что глаголы *искать* и *найти* очень тесно связаны синтагматически: они обладают высоким коэффициентом совместной встречаемости — как в прецедентных текстах (*кто ищет, тот всегда найдет*;

---

[3] Еще одна пара глаголов, которую следует упомянуть в рассматриваемой связи, это *искать — сыскать*. Глагол *сыскать*, имеющий ряд употребительных дериватов (ср. *сыск, сыскной, сыщик*), для современного русского языка является периферийным синонимом для *найти*; видовую пару он образует с еще более маргинальным, хотя и встречающимся в НКРЯ, глаголом *сыскивать*.

[4] Очевидно, это верно также для соответствующих пар глаголов в европейских языках: англ. *search* и *find,* нем. *suchen* и *finden*, фр. *chercher* и *trouver,* ит. *cercare* и *trovare*, исп. *buscar* и *trovar* и т. д. Идеи ‘искать’ и ‘найти’ обозначаются разными глаголами, которые при этом теснейшим образом семантически и синтагматически связаны.

бороться и искать, найти и не сдаваться и т. п.) так и в свободном дискурсе, ср. следующие примеры[5]:

> Ответ на этот вопрос он *искал* и *нашел* — на Земле, изучая движение с помощью двух своих главных инструментов — эксперимента и математически точного языка. [Геннадий Горелик. Гравитация — первая фундаментальная сила // «Знание-сила», 2012]

> Как Козинцев искал для «Короля Лира» «землю трагедии», так Тарковский *искал* — и *нашел* — для «Ностальгии» «воду трагедии» — замусоренный бассейн с дымящейся горячей водой, пахнущей серой. [Лидия Шодхина. Дорога к храму (2003) // «Вестник США», 2003.06.25]

Не менее характерна пара *искать* — *находить*[6], ср.:

> Авторы могучих сибирских романов *искали и даже находили горячий философский камень* [Денис Горелов. Москва кирзам верит. «Молодые». Режиссер Николай Москаленко. Год 1971. (2002) // «Известия», 2002.07.14]

> С тех пор я всегда чего нибудь *искал и часто находил неожиданные вещи*, так что прослыл чем-то вроде домашней ищейки. [Фазиль Искандер. Время счастливых находок (1973)]

> […] *ища и не находя глазами Нику*, я уже знал, что произошло. [Виктор Пелевин. Ника (1992)]

Анализ совместной встречаемости глаголов *искать* и *найти/находить* в НКРЯ дает впечатляющие результаты. Из 2090 случаев употребления последовательности *искать* плюс *и* (расстояние = 1) в 520 примерах далее следует глагол *найти* или *находить* (из них в 140 — с отрицанием), что составляет 24,9%. Другими словами, в четверти случаев употребления глагола *искать* в сопровождении союза *и* дальше следует глагол *найти* или *находить*[7].

Во-вторых, между глаголами со значением ‘искать’ и ‘найти’ в некоторых случаях возникает отношение квазисинонимии: вспомним компьютерные

---

[5] Здесь и далее — примеры из Национального корпуса русского языка (www.ruscorpora.ru).

[6] Напомним, что глагол НСВ *находить* является импефективным коррелятом к *найти*.

[7] Другой вариант поискового запроса — на встречаемость глаголов *найти/находить* на расстоянии 1–3 от глагола *искать* — дает 1,6% от всех вхождений этого глагола. Эта цифра на один-два порядка превосходит совместную встречаемость глагола *искать* с любым другим глаголом, а также вообще любую совместную встречаемость двух глаголов, связанных результативным отношением. Так, только 0,05% от всех вхождений глагола *думать* встречается в контексте глаголов *придумать/придумывать* на расстоянии 1–3.

функции *search* и *find*, по-русски *поиск* и *найти*: в обоих случаях речь идет о действии, целью которого является нечто *найти*.

Наконец, ‘искать’ и ‘найти’ иногда оказываются в отношении межъязыковой эквиваленции, ср. следующие примеры переводов:

> Изменив заведениям очевидным, где профан склонен был бы *искать* как раз его, он облюбовал это приличное, скучноватое кафе […]
> [В. В. Набоков. Весна в Фиальте (1938)]
> Having forsaken the two or three obvious haunts where naive amateurs of Montparnassian life would have expected to *find* him, he had started patronizing this perfectly bourgeois establishment […] [Vladimir Nabokov. Spring in Fialta (Peter Pertzov, Vladimir Nabokov, 1947)]

> Есть счастье, да нет ума *искать* его. [А. П. Чехов. Счастье (1887)]
> There is fortune, but there is not the wit to *find* it. [Anton Chekhov. Happiness (Constance Garnett, 1900–1930)][8]

> Не было сомнений, что и мы будем так жить и надо *искать* убежище. [Светлана Алексиевич. Время секонд хэнд (ч. 1) (2013)]
> Personne ne doutait une seconde que nous aussi, nous allions vivre ainsi, et qu’il fallait se *trouver* un refuge. [Svetlana Alexievitch. La fin de l’homme rouge ou le temps du désenchantement (p. 1) (Sophie Benech, 2013)].

2. Перейдем к ответу на второй вопрос — о соотношении между парой *искать — найти* и такими парами как *решать — решить* и *ловить — поймать*.

Прежде всего: действительно, между ними много общего, а именно, семантическое соотношение во всех трех парах — не собственно результативное, а конативное: желаемый результат здесь концептуализуется как не полностью контролируемый субъектом и, соответственно, по мере действия не обязательно происходит накопление результата; соответственно, глагол НСВ обозначает ‘попытку’, а глагол СВ — ‘успех’ в достижении этого результата.

Класс конативных глаголов несов. вида в русском языке включает глаголы *решать* <задачу>, *сдавать* <экзамен>, *убеждать*, *уговаривать*, *будить*, *соблазнять*, *поступать* <в университет>, *добиваться* <повышения в должности>, *зазывать* <кого-то в гости> и некоторые другие (ср. список в [Гловинская 2001: 104–106]). Однако далеко не все конативные пары являются бесспорно видовыми.

---

[8] Как справедливо заметил один из наших анонимных рецензентов, *нет ума искать* может обозначать ‘не додумывается искать’ — и в этом случае перевод является ошибочным. Однако в рассказе Чехова данной фразе предшествует следующий фрагмент, свидетельствующий о том, что переводчик правильно передал значение, которое здесь имелось в виду (речь идет о поисках клада): «— На своем веку я, признаться, раз десять искал счастья, — сказал старик, конфузливо почесываясь. — На настоящих местах искал, да, знать, попадал всё на заговоренные клады. И отец мой искал, и брат искал — ни шута не находили, так и умерли без счастья».

Напомним здесь масловское определение видовой пары, согласно которому критерием для объединения глаголов в пары служит не определенное семантическое соотношение между глаголом НСВ и СВ, а наличие контекстов, где семантическое различие между этими глаголами устраняется, а именно, когда глагол СВ заменяется на глагол НСВ без изменения значения:

> «Объективным критерием тождественности лексического значения префектива и имперфектива и основанием для их функционального объединения в составе одной лексемы служит возможность замены перфектива данным имперфективом. Такая замена возможна и даже обязательна в определенных типах контекстов, в частности — в историческом настоящем, в сценическом настоящем, при итеративизации». [Маслов 2004: 91]

Глаголы *решать* и *ловить*, имеющие значение ‘попытки’, в контекстах Маслова могут обозначать также и ‘успех’, т. е. принимать значение, равное значению глагола СВ, ср.: *решает одну задачу и принимается за другую; ловит бабочку и тут же выпускает ее на волю*. Глагол *искать* такой способностью не обладает и, тем самым, никак не может быть парой к *найти*[9].

Существенно, что, в отличие от собственно результативных видовых пар, где глаголы НСВ и СВ обычно переводятся на другие языки одним глаголом (ср.: *строить — построить*: англ. *to build*; *писать — написать*: англ. *to write*), перевод конативного глагола НСВ, если он употреблен в своем основном (не равном СВ) значении, требует экслицитного выражения идеи попытки. Тем самым, парный конативный глагол НСВ имеет два переводных эквивалента, ср.: *ловить* — англ. ‘try to catch’; ‘to catch’; *решать* — англ. ‘try to solve’; ‘to solve’.

При этом далеко не все конативные глаголы НСВ способны обозначать успешное достижение результата. Так, по крайней мере некоторые из них, включая те, которые связаны с глаголом СВ деривационными отношениями, типичными для парных глаголов, не проходят или с трудом проходят «тесты Маслова». Таковы, например, *умолять, убеждать, уговаривать, зазывать* <в гости>, ср.: *Он каждый раз умоляет/убеждает/уговаривает ее вернуться*: неизвестно, возвращается ли «она» каждый раз или нет. Для реализации событийного значения многим конативным глаголам нужен более «сильный» контекст, ср.: *Он* ***все-таки/наконец*** *уговаривает ее вернуться,* ***и*** *они отправляются в совместное путешествие*; ***Как только*** *он уговаривает ее вернуться, они отправляются в совместное путешествие*, и т. п. На самом деле

---

[9] Ср. следующее замечание Е. В. Падучевой: «Ситуацию X *ищет* Y русский язык (как, впрочем, и многие другие) не концептуализирует как деятельность с дискретной целью, которая обеспечила бы глаголу *искать* вхождение в предельную пару. Почему это так — особый вопрос. Ю. С. Маслов (1948) относит *искать* к глаголам «бесперспективного протекания». Ясно, что *искать* не относится к глаголам с накоплением результата: искомый предмет не становится все более найденным по мере деятельности ищущего — он вообще не является объектом воздействия» [Падучева 2004: 479].

событийное значение у конативного глагола почти всегда является «вынужденным», и разные конативные глаголы поддаются этому вынуждению в разной степени, ср.: ?*Как только он убеждает ее вернуться…*, ??*Как только он умоляет…*

Глагол *искать* этому вынуждению не поддается: нельзя сказать **Как только/наконец он ищет ключи, он открывает дверь и входит в квартиру*; **Он ищет ответ на этот мучительный вопрос и наконец успокаивается* — в этих контекстах необходимо употребить глагол *находить*.

Впрочем, существует класс контекстов, где *искать* имплицирует достижение результата ‘найти’ (так как последующее действие очевидно может осуществиться лишь после достижения данного результата), т. е. выступает в качестве квазисинонима глагола *находить*, ср. следующие примеры:

> Станислав Аксенов — экстремал. Он ***ищет*** высокие здания *и прыгает* с них. Так он тестирует городские высоты для своей команды the SINNER team… [Сайт // «Русский репортер», 2013]

> Когда ребенок тянется к винограду, помощник переводит его ручку на картинку, передает его рукой картинку психологам, а они сразу дают ему виноград. Так происходит не раз, пока ребенок не научится ***искать*** нужную карточку *и отдавать* ее человеку, у которого есть то, что он хочет. [Ольга Тимофеева. Выход № 1465 // «Русский репортер», 2015]

> Надо быстро ***искать*** деньги *и доплачивать* за трёшку. [Наши дети: Подростки (2004)]

> Брэнды в поисках новых рынков сбыта вынуждены вести себя точно так же: продвигать нестандартные идеи, ***искать*** сочетания несочетаемого *и создавать* новые привычки потребления. [Владимир Ляпоров. Молодая гвардия. Искусство быстрого завоевания новых рынков сбыта (2003) // «Бизнес-журнал», 2003.10.23]

> Теперь же ученые будут ***искать*** маркеры здоровья *и наблюдать*, под воздействием каких внешних факторов эти показатели повышаются или падают. [Елена Кудрявцева. Что съесть на завтра // «Огонек», 2014]

> Теперь Илья покупает не просто новые модели, а целенаправленно ***ищёт*** *и заказывает* те, которые хочет иметь в своём автопарке. [Елена Зиминова. Автопарк на столе (2012.11.24) // «Новгородские ведомости», 2012]

Действительно, во всех этих примерах описываются ситуации, когда действие *искать* приводит к результату ‘найти’. Однако, во-первых, компонент ‘найти’ здесь везде извлекается исключительно из контекста (т. е. является импликатурой) и, во-вторых, во всех этих примерах речь идет либо о хабитуальном, либо о потенциальном событии. Между тем для обозначения единичного

события глагол *искать* употреблен быть не может, а именно этот тип значения является критичным для вхождения глагола НСВ в видовую пару[10].

Почему же это так? Почему *искать* не может того, что может *ловить*, в чем между ними разница? Она состоит, очевидно, в том, что событие 'поймать' концептуализуется как в большей степени обусловленное предшествующим процессом, чем событие 'найти', которое, даже если ему предшествовали поиски, представлено языком как до какой-то степени случайное. В этом отношении *найти* приближается к *попасть* <в цель>: событие 'попасть' трактуется как не обусловленное предшествующим процессом и обозначается другим глаголом, хотя действие 'целиться' имеет очевидную цель — 'попасть'. Соответственно, *целиться* и *попасть* — это просто два разных, хотя и семантически связанных глагола. То же самое можно сказать и про *искать* и *найти*. Еще дальше от *ловить* отстоят, например, глаголы *преследовать* и *охотиться*, оба — типичные imperfectiva tantum, т. е. здесь вообще отсутствует какой-либо глагол, обозначающий результат, хотя оба действия имеют вполне определенную цель. Таким образом, ответ на второй вопрос состоит в том, что глагол *искать* находится в крайней точке шкалы степени предсказуемости наступления результата, на противоположном конце которой находятся образцовые предельные глаголы (ср. [Плунгян 2001: 61]).

3. Ответ на третий вопрос (чем пара *искать — найти* отличается от квазисинонимичных пар *разыскивать — разыскать* и *отыскивать — отыскать*) состоит в том, что глаголы *разыскать* и *отыскать* обозначают результат, обусловленный предшествующим контролируемым процессом поиска, ср. сомнительность ??*случайно разыскал/отыскал*. Между тем глагол *найти* относится к семантическому типу моментальных глаголов и сам по себе никакого действия не предполагает. Возможно, именно поэтому так частотна совместная встречаемость глаголов *найти* и *искать*: в случаях, когда событие 'найти' наступило не случайно, а в результате поиска, это обстоятельство должно быть выражено эксплицитно[11]. С другой стороны, глаголы *разыскивать* и *отыскивать* — это вторичные имперфективы, для которых наличие событийного значения — это норма: вторичные имперфективы обычно достаточно точно воспроизводят семантику префиксального глагола сов. вида, от которого они произошли; объяснения требуют, наоборот, случаи отсутствия у таких глаголов событийного значения (ср. выше о глаголах *умолять, уговаривать* и т. п.). Тем

---

[10] В книге [Гловинская 2001: 107] на основании контекстов, где глагол *искать* имеет результативное употребление в потенциальном значении (типа *Эта собака хорошо ищет наркотики*) предлагается считать *искать* членом видовой пары — с чем мы, однако, не можем согласиться.

[11] Один из наших анонимных рецензентов заметил, что глагол *разыскивать*, в отличие от глагола *искать*, включает презумпцию существования объекта поиска. Действительно, можно сказать *Он разыскивает свою жену* и нельзя **Он разыскивает себе жену* — при том, что можно: *Он ищет себе жену.* Очевидно, с этим связано и то, что объектом глагола *разыскивать* не может быть абстрактная сущность, ср. **Он разыскивает смысл жизни*.

самым, *разыскивать — разыскать* и *отыскивать — отыскать* это полноценные видовые пары с конативным семантическим соотношением.

Приведем несколько примеров, где глаголы *разыскивать* и *отыскивать* имеют событийное значение — соответственно, ‘разыскать’ и ‘отыскать’. Заметим, что в каждом из них эти глаголы могут быть заменены на глагол *находить*, но с очевидной смысловой потерей (а именно, с уменьшением степени контролируемости результата). Однако ни в одном из этих примеров не может быть использован глагол *искать*.

> Он занялся также астрономическими определениями главных пунктов, через которые пролегал его путь, обучая практически астрономии и геодезии сопровождавших его учеников, *и, наконец, разыскивал* разные специальные старые русские карты в архивах сибирских городов. [Б. Г. Островский. Великая Северная экспедиция (1937)]

> Джек узнает правду о жизни его отца, влюбляется в сводную сестру *и, наконец, разыскивает* таинственную Натали. [http://ratings.7ya.ru/books/Indigo]

> Зато корреспондент НТВ Елена Масюк (а до нее — журналисты Франс Пресс) *легко отыскивает* Басаева и берет у него интервью, причем точно указывает в своем репортаже место, где произошла их встреча. [Игорь Малашенко. Заговор кукол. А куклы кто? (1995) // «Общая газета», 1995.07.26]

> Фауст смутно стремится к чему-то, чего-то ищет, бросается от наслаждения к наслаждению *и наконец отыскивает* смысл жизни, *отыскивает* свое счастье. [http://www.monsalvat.globalfolio.net/frglorios/parcifal/oparcifale.htm]

В заключение отметим, что наличие в русском языке видовых пар *разыскивать — разыскать* и *отыскивать — отыскать* фактически означает, что русский язык умеет выражать концепты ‘искать’ и ‘найти’ одним словом, — что, однако, никак не влияет на аспектуальный статус самой пары *искать — найти*. Отсутствие у глагола *искать* событийного значения, которое обеспечивало бы ему вхождение в видовую пару, имеет вполне отчетливые семантические основания, которые обнаруживаются при сравнении *искать* с такими конативными глаголами как *решать* или *ловить*, а также с предельными глаголами *разыскивать* и *отыскивать*.

Мы пользуемся возможностью выразить благодарность нашим анонимным рецензентам, чьи замечания мы постарались учесть в окончательной версии статьи.

Литература

1. *Апресян Ю. Д.* (1974/95), Лексическая семантика, М., 1974. (Ю. Д. Апресян. Избранные труды. Т.1. М., 1995).
2. *Апресян Ю. Д.* (1980), Типы информации для поверхностно-семантического компонента модели «Смысл<==>Текст» // Wiener slavistischer Almanach. Sbd. 1. Wien, 1980 (или: *// Апресян Ю. Д.* Избранные труды. Т. II. М., 1995. С. 8–101).
3. *Булыгина Т. В., Шмелев А. Д.* (1997), Типы каузации и лексикографическое описание русских каузативов // Т. В. Булыгина, А. Д. Шмелев. Языковая концептуализация мира (на материале русской раматики). Москва.
4. *Гловинская М. Я.* (1982), Семантические типы видовых противопоставлений русского глагола. Москва.
5. *Гловинская М. Я.* (2001), Многозначность и синонимия в видо-временной системе русского глагола. Москва.
6. *Зализняк Анна А., Шмелев А. Д.* (2000), Введение в русскую аспектологию. М., 2000.
7. *Исаченко А. В.* (1960) Грамматический строй русского языка в сопоставлении со словацким. Т. II. Братислава.
8. *Маслов Ю. С.*(1984/2004), Вид и лексическое значение глагола // Ю. С. Маслов. Избранные труды. Аспектология. Общее языкознание. Москва.
9. *Озеров Н. Н.* (2006), Краткий очерк синтаксиса русского языка. Простое предложение. — Москва: Языки славянской культуры.
10. *Падучева Е. В.* (2004), Динамические модели в семантике лексики. — Москва: Языки славянской культуры.
11. *Плунгян В. А.* (2001), Антирезультатив: до и после результата. // Исследования по теории грамматики. Вып. 1. Глагольные категории. Москва: Русские словари. С. 50–88.
12. *Шмелева Т. В.* (2011), Современный русский язык: морфология. Учебное пособие по циклу практических и семинарских занятий. Красноярск.
13. *Шустикова Т. В.* (ред.) (2012), Русские глаголы. Формы и контекстное употребление: учебное пособие / под ред. проф. Т. В. Шустиковой, Флинта, Москва.

Abstracts

THE IMPACT OF DIFFERENT DATA SOURCES ON FINDING AND RANKING SYNONYMS FOR A LARGE-SCALE VOCABULARY

Antonova A. (antonova@yandex-team.ru), **Kobernik T.** (kobernik@yandex-team.ru), **Misyurev A.** (misyurev@yandex-team.ru), Yandex, Moscow, Russia

In this paper we compare different models for measuring synonymy. We consider methods based on monolingual text corpora and parallel texts. We experiment with the features based on context similarity, translation similarity, and similarity of neighbors in the parse trees. We provide an analysis of strong and weak points of different approaches and show that their combination can improve the results. The considered methods can handle large-scale vocabularies and be useful for automatic construction of human-oriented synonym dictionaries.

ISCHEZNUT' 'TO DISAPPEAR' AND *PROPAST'* 'TO VANISH': POLYSEMY AND SEMANTIC MOTIVATION

Apresjan V. Ju. (valentina.apresjan@gmail.com), National Research University Higher School of Economics; Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

The paper considers Russian synonymic verbs *ischeznut'* 'to disappear' and *propast'* 'to vanish' and analyzes their semantic differences which motivate their differences in syntactic, aspectual and collocational properties, as well as in their polysemies. Semantic oppositions that distinguish between these two verbs, namely, type and referential status of the disappearing object, cause of disappearance, speaker's expectations, speed and completeness of disappearance, presence of an observer can be applied to the analysis of the entire semantic domain of the 'end of existence'.

SEMANTICS AND PRAGMATICS OF THE RUSSIAN WORDS *POSLEDNIJ* AND *PREDPOSLEDNIJ*

Apresjan V. Ju. (valentina.apresjan@gmail.com), National Research University Higher School of Economics; Vinogradov Russian Language Institute of the Russian Academyof Sciences, Moscow, Russia; **Shmelev A. D.** (shmelev.alexei@gmail.com), Moscow Pedagogical State University; Vinogradov Institute of Russian Language, Russian Academy of Sciences; St Tikhon's Orthodox University, Moscow, Russia

The paper considers the senses of the Russian adjective *poslednij* 'last'. Its polysemy is analyzed as deriving from a certain core semantic structure that is common to all its meanings. The core structure has two semantic valencies — of a sequence and of a sequence element. Modifications of the core structure, including additional valencies (point of reference and landmark) account for its polysemy, as well as for diversity of its collocational and syntactic properties. The paper also demonstrates the role of pragmatics and lexicalization of grammatical and syntactic forms in disambiguating different meanings of *poslednij*, against the backdrop of its English correlates.

DEVELOPING A POLYSYNTHETIC LANGUAGE CORPUS: PROBLEMS AND SOLUTIONS

Arkhangelskiy T. A. (tarkhangelskiy@hse.ru), **Lander Yu. A.** (yulander@hse.ru), National Research University Higher School of Economics, Moscow, Russia

Although there exist comprehensive morphologically annotated corpora for many morphologically rich languages, there have been no such corpora for any polysynthetic language so far. Developing a corpus of a polysynthetic language poses a range of theoretical and practical challenges for corpus linguistics. Some of these challenges have been partly addressed when developing corpora for languages with extensive morphological inventories and numerous pro-

ductive derivation models such as Turkic or Uralic, while others are unique for this kind of languages. As we are currently working on a corpus of the polysynthetic West Circassian language, we had to identify these challenges and propose theoretical and practical solutions. These include the tokenization problem, which involves delimiting morphology from syntax, the problem with lemmatization and part-of-speech tagging, and a number of glossing and search issues. The solutions proposed in the paper are partly implemented and will be available for public testing when the preliminary version of the corpus is released.

COMPARISON OF NEURAL NETWORK ARCHITECTURES FOR SENTIMENT ANALYSIS OF RUSSIAN TWEETS

Arkhipenko K. (arkhipenko@ispras.ru)[1,2], **Kozlov I.** (kozlov-ilya@ispras.ru)[1,3], **Trofimovich J.** (integral@ispras.ru)[1], **Skorniakov K.** (kirill.skorniakov@ispras.ru)[1,3], **Gomzin A.** (gomzin@ispras.ru)[1,2], **Turdakov D.** (turdakov@ispras.ru)[1,2,4]
[1]Institute for System Programming of RAS, Moscow, Russia; [2]Lomonosov Moscow State University, CMC faculty, Moscow, Russia; [3]MIPT, Dolgoprudny, Russia; [4]FCS NRU HSE, Moscow, Russia

The paper presents evaluation of three neural network based approaches to Twitter sentiment analysis task performed at SentiRuEval-2016. The task focuses on sentiment classification of tweets about banks and telecommunication companies.

Our team submitted three solutions which are based on different supervised classifiers: Gated Recurrent Unit neural network (GRU), convolutional neural network (CNN), and SVM classifier with domain adaptation combined with previous two classifiers. We used vector representations of words obtained with word2vec model as features for classifiers. These classifiers were trained on labeled data provided by organizers of the evaluation. Additionally, we collected several million posts and comments from social networks for training word2vec model. According to evaluation results, GRU-based solution shows the best macro-averaged F1-score for both domains (banks and telecommunication companies) and also has the best micro-averaged F1-score for banks domain among all solutions submitted to SentiRuEval.

LINGUISTIC DISFLUENCY IN CHILDREN DISCOURSE: LANGUAGE LIMITATIONS OR EXECUTIVE STRATEGY?

Balčiūnienė I. (i.balciuniene@hmf.vdu.lt)[1,2], **Kornev A. N.** (k1949@yandex.ru)[1]
[1]Vytautas Magnus University, Kaunas, Lithuania
[2]Saint-Petersburg State Pediatric Medical University, Saint-Petersburg, Russia

The paper deals with linguistic disfluencies (hesitations, repetitions, revisions, false starts, and incomplete utterances) in Russian-speaking language-impaired (N=12) vs. typically-developing (N=12) preschoolers. The corpus-based study aimed at evaluation and comparison of linguistic disfluency in narrative vs. dialogue discourse within and between the groups. Following the *Russian Assessment Instrument for Narrative* (RAIN) methodology, each subject performed two tasks, i.e. storytelling and story retelling according wordless picture sequences; each of the tasks was followed by a structured dialogue based on ten comprehension questions. Both narratives and dialogues were transcribed and annotated for automatized linguistic analysis. Finally, individual measures (a number of each category of disfluencies per utterance) were estimated and submitted for statistical analysis.

Results of our study evidenced that mainly linguistic disfluencies are caused by distinct strategies of speech production due to a level of the subject's language competence, cognitive resource, and the circumstances of narrative and dialogue production.

ABOUT DISCURSIVE MODES OF EVALUATION IN RUSSIAN

Baranov A. N. (baranov_anatoly@hotmail.com), Institute of Russian Language, National Research University Higher School of Economics, Moscow, Russia

The paper discusses different modes of evaluation in Russian. Evaluation is considered as a speech act based on a cognitive procedure which has the following form: (i) evaluation of an object X as possessing a feature q consists of comparing of parameter Q with X and picking out of

q as a function of *Q* with an argument of *X*; (ii) the feature *q* presupposes recommendations for decision making in connection with an object *X*. Cognitive procedure of description of an object X as possessing of a feature *q* doesn't presupposes any recommendation for decision making.

In some discursive modes semantics of evaluation lose its influence force or at least it getting weaker. Discursive mode is defined as a sphere of functioning of speech forms in discourse, in which their meaning regularly changed. Different discourses allow different kinds of discursive modes. In the paper are discussed the following discursive modes, which modify evaluation force: irony, language game, common nomination, indefinite reference.

VERY LARGE RUSSIAN CORPORA: NEW OPPORTUNITIES AND NEW CHALLENGES

Benko V. (vladob@juls.savba.sk), Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics, Bratislava, Slovakia; **Zakharov V. P.** (v.zakharov@spbu.ru), St. Petersburg State University; Institute for Linguistic Studies, RAS, St. Petersburg, Russia

Our paper deals with the rapidly developing area of corpus linguistics referred to as *Web as Corpus* (*WaC*), i.e., creation of very large corpora composed of texts downloaded from the web. Some problems of compilation and usage of such corpora are addressed, most notably the "language quality" of web texts and the inadequate balance of web corpora, with the latter being an obstacle both for corpus creators, and its users. We introduce the *Aranea* family of web corpora, describe the various processing procedures used during its compilation, and present an attempt to increase the size of its Russian component by the order of magnitude. We also compare its contents from the user's perspective among the various sizes of the Russian *Aranea*, as well as with the other large Russian corpora (*RNC*, *ruTenTen* and *GICR*). We also intent to demonstrate the advantage of a very large corpus in linguistic analysis of low-frequency language phenomena in linguistics, such as usage of idioms and other types of fixed expressions.

THE BEGINNING OF A BEAUTIFUL FRIENDSHIP: RULE-BASED AND STATISTICAL ANALYSIS OF MIDDLE RUSSIAN

Berdičevskis A. (aleksandrs.berdicevskis@uit.no), **Eckhoff H.** (hanne.m.eckhoff@uit.no), UiT The Arctic University of Norway, Tromsø, Norway; **Gavrilova T.** (tanya96gavrilova@gmail.com), The National Research University "Higher School of Economics", Moscow, Russia

We describe and compare two tools for processing Middle Russian texts. Both tools provide lemmatization, part-of-speech and morphological annotation. One ("RNC") was developed for annotating texts in the Russian National Corpus and is rule-based. The other one ("TOROT") is being used for annotating the eponymous corpus and is statistical. We apply the two analyzers to the same Middle Russian text and then compare their outputs with high-quality manual annotation. Since the analyzers use different annotation schemes and spelling principles, we have to harmonize their outputs before we can compare them. The comparison shows that TOROT performs considerably better than RNC (lemmatization 69.8% vs. 47.3%, part of speech 89.5% vs. 54.2%, morphology 81.5% vs. 16.7%). If, however, we limit the evaluation set only to those tokens for which the analyzers provide a guess and in addition consider the RNC response correct if one of the multiple guesses it provides is correct, the numbers become comparable (88.5% vs. 91.9%, 93.9% vs. 95.2%, 81.5% vs. 86.8%). We develop a simple procedure which boosts TOROT lemmatization accuracy by 8.7% by using RNC lemma guesses when TOROT fails to provide one and matching them against the existing TOROT lemma database. We conclude that a statistical analyzer (trained on a large material) can deal with non-standardised historical texts better than a rule-based one. Still, it is possible to make the analyzers collaborate, boosting the performance of the superior one.

ESTIMATING SYNTAGMATIC ASSOCIATION STRENGTH USING DISTRIBUTIONAL WORD REPRESENTATIONS

Bukia G. T. (gregorybookia@yandex.ru), **Protopopova E. V.** (protoev@yandex.ru), **Panicheva P. V.** (p.panicheva@spbu.ru), **Mitrofanova O. A.** (o.mitrofanova@spbu.ru), St. Petersburg State University, St. Petersburg, Russia

Abstract: In the paper we present distributed vector space models based on word embeddings and a specific association-oriented count-based distributional algorithm which have been applied to measuring association strength in Russian syntagmatic relations (namely, between nouns and adjectives). We discuss the compositional properties of the vectors representing

nouns, adjectives and adjective-noun compositions and propose two methods of detecting the syntactic association possibility. The accuracy of the proposed measures is evaluated by means of a pseudo-disambiguation test procedure and all models show considerably high results. The errors are manually annotated, and the model errors are classified in terms of their linguistic nature and compositionality features.

USING CONSTRAINTS ON A GENERAL KNOWLEDGE LEXICAL NETWORK FOR DOMAIN-SPECIFIC SEMANTIC RELATION EXTRACTION AND MODELING

Clairet N. (clairet@lirmm.fr)[1,2,3], **Ramadier L.** (ramadier@lirmm.fr)[1,4], **Lafourcade M.** (mathieu@lafourcade@lirmm.fr)[1]
[1]Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), Montpellier, France; [2]Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), Paris France; [3]Lingua et Machina, Boulogne-Billancourt, France; [4]IMAIOS, 672, Montpellier, France

We introduce a pattern-based approach applied to the semantic relation retrieval and semantic modeling. Our method relies upon the use of a general knowledge lexical semantic network built, shaped, and handled by crowdsourcing and GWAPs (games with a purpose). Implementing constraints on semantic relations available in the network increases the efficiency of the relation extraction process but also opens a semantic modeling perspective. In terms of (mostly horizontal) relation extraction, we tested our method on radiology reports in French. Our results show the interest of using a general knowledge lexical semantic network for the domain specific textual analysis as well as the interest of implementing series of constraints on semantic relations for the relation retrieval. We recently turned to the analysis of cooking recipes that stand for examples of domain specific instructional texts. Thus, in addition to the semantic relation discovery, we are building a method for the semantic modeling and conceptualization of cooking instructions. Its first results are presented below. Today, our results are available for French but we target extending the lexical network coverage to other languages in the next few years.

THE DISCURSIVE CONSTRUCTION *ДЕЛО В ТОМ, ЧТО* AND ITS PARALLELS IN OTHER LANGUAGES: A CONTRASTIVE CORPUS STUDY

Dobrovol'skij D. (dm-dbrv@yandex.ru), Russian Language Institute, Russian Academy of Sciences, Moscow, Russia; **Pöppel L.** (ludmila.poppel@slav.su.se), Stockholm University, Department of Slavic and Baltic languages, Finnish, Dutch and German, Stockholm, Sweden

The primary goal of the present study is to improve methods for contrastive corpus investigations. Our data is the Russian construction *дело в том, что* and its parallels in English, German and Swedish. This construction, which appears to present no difficulty for translation into other languages, is in fact language-specific with respect to at least one parameter. It displays a large number of different parallels (translation equivalents) in other languages, and possesses a complex semantic structure. The configuration of semantic elements comprising the content plane of this construction is unique. The empirical data have been collected from the corpus query system Sketch Engine, subcorpus OPUS2 Russian, and the Russian National Corpus (RNC). The analysis shows that the construction *дело в том, что* has more than 50 parallels in English, over 30 in German, and about 30 in Swedish. In all three languages the most common means of translating the construction is to omit it. Also frequent are the English equivalents *the fact/thing/point/truth is (that)*; *(it's/this/that is) because*; the German expressions *nämlich*; *die Sache ist, die*; *denn*; and the Swedish constructions *saken är den att*; *problemet/faktum är att*. The semantic structure of *дело в том, что* includes the following components: 1) substantiation of something stated previously; 2) indication of the reason something has happened; 3) emphasis on the significance of what has been stated. The different translations of the construction are motivated by the fact that each specific context focuses on one of these meanings.

AUTOMATIC GENERATION OF THE DOMAIN-SPECIFIC SENTIMENT RUSSIAN DICTIONARIES

Dubatovka A. (alina.dubatovka@gmail.com), Saint Petersburg State University, St. Petersburg, Russia; **Kurochkin Yu.** (yurakura@yandex-team.ru), Yandex, St. Petersburg, Russia; **Mikhailova E.** (e.mikhaylova@spbu.ru), Saint Petersburg State University, St. Petersburg, Russia

This paper presents an algorithm for generating the Domain-Specific Sentiment Russian dictionary using a graph model. It is important to emphasize that the described algorithm does not require any human-labeling, but just a sufficiently large corpus of Russian texts from the subject area, which can be generated automatically for most domains. Our algorithm is not strictly confined to the Russian language and, if necessary, can be generalized to develop dictionaries in other languages.

Dictionaries of positive and negative words are created using the analysis of the graph constructed on unlabeled corpus of the Domain-Specific Russian texts. The graph was built using the approach described in [6], pre-adapted to texts in Russian. The applicability of this method to create a graph for prediction of polarity of adjectives in reviews in Russian language is experimentally evaluated.

The original method of graph processing for splitting the vertex set of this graph into subsets of positive and negative words was proposed and implemented. The algorithm starts with gathering a small seed set of adjectives, polarity of which is unambiguous irrespective of a subject area (for example, “bad”, “good”, “terrible”, “excellent”).

Further, words are distributed iteratively: each time a vertex is added to the set, if the vertex is most strongly associated with the already existing vertices in the set. Several weighting functions on the edges were compared, as well as functions of attraction to the sets of positive and negative words with the aim of composing the most accurate dictionaries of positive and negative adjectives for a specific subject area.

TEMPORAL COORDINATION BETWEEN GESTURAL AND SPEECH UNITS IN MULTIMODAL COMMUNICATION

Fedorova O. V. (olga.fedorova@msu.ru)[1,2,3], **Kibrik A. A.** (aakibrik@gmail.com)[1,2], **Korotaev N. A.** (n_korotaev@hotmail.com)[2,3,4], **Litvinenko A. O.** (allal1978@gmail.com)[2], **Nikolaeva Ju. V.** (julianikk@gmail.com)[1,2], [1]Lomonosov Moscow State University, Moscow, Russia; [2]Institute of Linguistics RAS, Moscow, Russia; [3]RANEPA, Moscow, Russia; [4]RSUH, Moscow, Russia

This study contributes to the research field of multimodal linguistics. Multimodal linguistics explores numerous channels involved in natural communication, such as verbal structure, prosody, gesticulation, mimics, eye gaze, etc., and treats them as parts of an integral process. Among the key issues in multimodal studies is the question of temporal coordination between the illustrative manual gestures (that is, spontaneous co-speech gestures) and elementary discourse units (that is, basic quanta of the local structure of spoken discourse). We address this issue with the help of a novel multimodal corpus “Pear chats and stories” that is currently under construction. It had been shown in a number of studies that gesture onset usually precedes speech onset. In order to verify this claim through our materials, we developed an analytic method that allowed to conduct a more detailed study. According to our results, it is only less than a half of all gestures that are produced before the corresponding fragment of talk. The most likely explanation of the obtained results is associated with gestures’ affiliation in a certain functional class, that is strongly dependent on discourse genre and speakers’ individual differences.

STYLE AND GENRE CLASSIFICATION BY MEANS OF DEEP TEXTUAL PARSING

Galitsky B. A. (bgalitsky@hotmail.com), **Ilvovsky D. A.** (dilvovsky@hse.ru), **Chernyak E. L.** (echernyak@hse.ru), **Kuznetsov S. O.** (skuznetsov@hse.ru), National Research University Higher School of Economics, Moscow, Russia

In this paper we show that using deep textual parsing, which is finding complex features such as syntactic and discourse structures of the text, helps to improve the quality of style and genre classification. These results confirm achievements of many researches that have many times stated that using syntactic or morphological pattern for style and genre classification results in poor precision

and recall. The best practice so far is to use n-gram patterns for this type of text classification problem. Syntactic and discourse structures allow however to capture some style of genre specific pattern of texts and to reach average precision higher than 95% on binary multi-genre classification.

RUSSIAN VERBAL ASPECT AND GESTICULATION

Grishina E. A. (rudi2007@yandex.ru), Vinogradov Institute of Russian Language RAS, Moscow, Russia

In this paper, we use evidence from the Multimodal Russian Corpus (MURCO) to explore gesture properties that enable distinction between perfective and imperfective Russian verbs. The properties identified are duration, repetition, and energy. We show that repetition and energy differentiate perfective and imperfective verbs because these properties are salient in gestures accompanying one group of verbs but are not manifest in gestures accompanying verbs in the other group. Gesture duration, on the other hand, can be used to identify either aspect.

THE STRUCTURE OF TWO-PART CORRELATIVE CONNECTORS AS AN OBJECT OF CORPUS ANALYSIS

Inkova O. Yu. (Olga.Inkova@unige.ch), University of Geneva, Geneva, Switzerland; Institute of Informatics Problems, FRC CSC RAS, Moscow, Russia; **Popkova N. A.** (Natasha__popkova@mail.ru), Institute of Informatics Problems, FRC CSC RAS, Moscow, Russia

The paper discusses the problem of formal variability of Russian two-part correlative connectors on the example of *ne to chtoby…no* and *ne to chto…a*. The results of the analysis, carried out both with formal and functional-semantic criteria, allow to state that *ne to chtoby… no* and *ne to chto… a* are two separate linguistic units with the first expressing substitution aimed towards more descriptive adequacy and the second unit expressing, beyond that, substitution aimed towards more argumentative relevance. This semantic difference is due to the different scope of the negative particle *ne*, which is the part of both markers; even if in both cases the gradation is rising. The position of *ne to chtoby* and *ne to chto* is not fixed, and *ne to chtoby* can be characterized by phonetic (*ne to chtob*) and morphological (*ne tak chtob(y)*) variability. As forms *ne to chtoby* and *ne to chto* can express relations of substitution alone, they may be considered basic or minimal markers of such relations. The use of these forms as two-part correlative connectors with adversative conjunctions *no, a* and other lexical units is dictated by the speaker's communicative intention, the syntactical construction and other discursive parameters. The Russian National Corpus data confirms our statements.

WORD SENSE FREQUENCY OF SIMILAR POLYSEMOUS WORDS IN DIFFERENT LANGUAGES

Iomdin B. L. (iomdin@ruslang.ru), V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences; National Research University "Higher School of Economics", Moscow, Russia; **Lopukhin K. A.** (kostia.lopuhin@gmail.com), Scrapinghub, Moscow, Russia; **Lopukhina A. A.** (nastya-merk@yandex.ru), V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia; **Nosyrev G. V.** (grigorij-nosyrev@yandex.ru), Yandex, Moscow, Russia

When words have several senses, it is important to describe them properly in dictionary (a lexicographic task) and to be able to distinguish them in a given context (a computational linguistics task, WSD). Different senses normally have different frequencies in corpora. We introduced several techniques for determining sense frequency based on dictionary entries matched with data from large corpora. Information about word sense frequency is not only useful for explanatory lexicography and WSD, but it also may enrich language learning resources. Learners of a foreign language who encounter a word similar to one of their native language are often tempted to assume that the foreign word and its equivalent have the same meaning structure. Sometimes, however, this is not the case, and the most frequent sense of a word in one language may be much less frequent for its cognate. We proposed a method for detecting such cases. Having selected a set of Russian words included into the Active Dictionary of Russian which have more than two dictionary senses and have cognates in English, we estimated the frequencies for English and Russian senses using SemCor and Russian National Corpus respectively, matched the senses in each pair of words and compared their frequencies. Thus we revealed cases in which the most frequent senses

and whole meaning structures are, cross-linguistically, substantially different and studied them in more detail. This technique can be applied not only to cognates, but also to pairs of words which are usually offered by the dictionaries as the translation equivalents of each other.

ENTITY BASED SENTIMENT ANALYSIS USING SYNTAX PATTERNS AND CONVOLUTIONAL NEURAL NETWORK

Karpov I. A. (karpovilia@gmail.com)[1], **Kozhevnikov M. V.** (kozhevnikov1511@gmail.com)[2], **Kazorin V. I.** (zhelyazik@mail.ru)[2], **Nemov N. R.** (nemo_1@pisem.net)[2]
[1]National Research University Higher School of Economics, Moscow, Russia
[2]Research and Development Institute «Kvant», Moscow, Russia

This paper provides an alternative method to extracting object-based sentiment in text messages, based on modified method previously proposed by Mingbo [8], in which we first parse the syntax, and then correlate the sentiment with the object of analysis (also referred to as entity by some, therefore, used in this article interchangeably). We show two approaches for the sentiment polarity classification: syntactic rule patterns and convolutional neural network (CNN). Even without domain specific vocabulary and sophisticated classification algorithms, rule-based approach demonstrates an average macro-F_1 based rank among the participants, whereas domain-specific vocabularies show a slightly higher macro-F_1 score, but still close to an average result. CNN approach uses syntax dependencies and linear word order to obtain more extensive information about object relations. Convolution patterns, designed in this approach, are very similar to rules, obtained with rule-based approach. In our proposed approach, the neural network was trained with different Word2Vec (WV) models; we compared their performance relative to each other. In this paper, we show that learning a domain-specific WV offers slight progress in performance. Resulting macro-F_1 score show performance in the into top three of the overall results among the competitors, participating in 2016 SentiRuEval event. Originally, we have not submitted our results to this competition at the time it was held, but had a chance to compare them post-hoc. We also combine the CNN approach with the rule-based approach and discuss the obtained differences in results. All training sets, evaluation metrics and experiments are used according to SentiRuEval 2016.

LARGE CORPORA AND FREQUENCY NOUNS

Khokhlova M. V. (m.khokhlova@spbu.ru), St. Petersburg State University, St. Petersburg, Russia

The paper describes a new branch in corpus linguistics that deals with building and using large corpora. We introduce several new large Russian corpora that have recently become available. The paper gives a survey of the given corpora and analyzes a number of Russian nouns across the following corpora of different sizes: the Russian Web corpus by S. Sharoff (187.97 mln tokens), ruTenTen (18.28 bln tokens) and its sample (1.25 bln tokens). The research focuses on the discussion on these corpora, their comparison and the study of frequency properties for the high- and low frequency Russian nouns comparing them with data published in the Frequency Dictionary. The analysis shows the lists presented in the frequency dictionary of Russian differs from the corpus data depending on types of the nouns.

VOICE COARTICULATION ACROSS WORD BOUNDARIES AS A CUE FOR DETECTING PROSODIC BREAKS IN STANDARD RUSSIAN

Knyazev S. V. (svknia@gmail.com), Moscow State University, Higher School of Economics; Moscow, Russia

The paper reports some results of the research, aimed at finding out whether regressive and / or progressive voice coarticulation available in clusters of homorganic labiodental consonants /v/ + /v/ in an external sandhi position in Modern Standard Russian may serve as a cue for detecting the location and depth of prosodic breaks.

Combinations of labiodental fricatives /v/ + /v/ at the word junctures result in [ff], [vv] or [fv] pronunciation (with the decreasing abundance) in Modern Standard Russian. The percentage ratio of the above mentioned pronunciation types depends on the strength of the prosodic break between two words:

- in the position within an intonation group (no prosodic break) [ff] pronunciation appears fairly stable and makes about 70% of the total case number, while the percentage of [fv] pronunciation (corresponding to the absence of the coarticulation) varies in the range of 1–11%.
- in the position around prosodic break between two words group [fv] pronunciation detected in more then 80% out of the total case number studied.

"AS IF IT WASN'T ME AND IF IT WASN'T YOU": PRAGMATICS OF REFERENTIAL GLIDE IN SPOKEN LANGUAGE

Kolmogorova A. V. (nastiakol@mail.ru), Siberian Federal University, Krasnoyarsk, Russia

The article explores pragmatic functions of the "communicative mummery"—phenomenon observable in everyday communication and consisting in some referential slides that happen in the communicative act trivial schema "I talk to You Here and Now": the speaker can communicate as if he wasn't himself but someone else or if he was talking to another person but not to his real communicative partner. Unlike a simple trickery the communicative mummery isn't hidden from the speaker's interlocutor. Conversely, experiencing such transfigurations (I speak, if it wasn't me…etc.) the speaker intentionally uses a range of prosodic and/or non verbal markers such as special gaze, gesturing that isn't familiar to him, specific accent or prosodic contours for attracting his partner attention. The discursive manifestations of communicative mummery have some common features with the reported speech and the polyphonic conversational humor phenomena but, in the same time, display its own particulars properties and perform rather special functions in conversation. Firstly remarked in mother-child communication as a particular mother's practice of child socialization the discussed phenomenon was also found in adults' heterogeneous speech interactions that, after having been collected in a corpus of 52 items, served as a data for our analysis. It shows that the main pragmatic functions of communicative mummery is to prevent the speaker's social face loss in case if he violates social conventions regulating communicative behavior in the situations of social enforcement, social guilt or self-praising.

AN OPINION WORD LEXICON AND A TRAINING DATASET FOR RUSSIAN SENTIMENT ANALYSIS OF SOCIAL MEDIA

Koltsova O. Yu. (ekoltsova@hse.ru)[1], **Alexeeva S. V.** (salexeeva@hse.ru)[1, 2], **Kolcov S. N.** (skoltsov@hse.ru)[1]
[1]National Research Institute Higher School of Economics, St. Petersburg, Russia
[2]St. Petersburg State University, St. Petersburg, Russia

Automatic assessment of sentiment in large text corpora is an important goal in social sciences. This paper describes a methodology and the results of the development of a system for Russian language sentiment analysis that includes: a publicly available sentiment lexicon, a publicly available test collection with sentiment markup and a crowdsourcing website for such markup. The lexicon is aimed at detecting sentiment in user-generated content (blogs, social media) related to social and political issues. Its prototype was formed based on other dictionaries and on the topic modeling performed on a large collection of blog posts. Topic modeling revealed relevant (social and political) topics and as a result—relevant words for the lexicon prototype and relevant texts for the training collection. Each word was assessed by at least three volunteers in the context of three different texts where the word occurred while the texts received their sentiment scores from the same volunteers as well. Both texts and words were scored from −2 (negative) to +2 (positive). Of 7,546 candidate words, 2,753 got non-neutral sentiment scores. The quality of the lexicon was assessed with *SentiStrength* software by comparing human text scores with the scores obtained automatically based on the created lexicon. 93% of texts were classified correctly at the error level of ±1 class, which closely matches the result of *SentiStrength* initial application to the English language tweets. Negative classes were much larger and better predicted. The lexicon and the text collection are publicly available at http://linis-crowd.org.

IMPROVING DISTRIBUTIONAL SEMANTIC MODELS USING ANAPHORA RESOLUTION DURING LINGUISTIC PREPROCESSING

Koslowa O. (evezhier@gmail.com), National Research University Higher School of Economics, Moscow, Russia; **Kutuzov A.** (andreku@ifi.uio.no), University of Oslo, Norway

In natural language processing, distributional semantic models are known as an efficient data driven approach to word and text representation, which allows computing meaning directly from large text corpora into word embeddings in a vector space. This paper addresses the role of linguistic preprocessing in enhancing performance of distributional models, and particularly studies pronominal anaphora resolution as a way to exploit more co-occurrence data without directly increasing the size of the training corpus.

We replace three different types of anaphoric pronouns with their antecedents in the training corpus and evaluate the extent to which this affects the performance of the resulting models in lexical similarity tasks. CBOW and SkipGram distributed models trained on Russian National

Corpus are in the focus of our research, although the results are potentially applicable to other distributional semantic frameworks and languages as well. The trained models are evaluated against RUSSE'15 and SimLex-999 gold standard data sets. As a result, we find that models trained on corpora with pronominal anaphora resolved perform significantly better than their counterparts trained on baseline corpora.

MANUALLY CREATED SENTIMENT LEXICONS: RESEARCH AND DEVELOPMENT

Kotelnikov E. V. (kotelnikov.ev@gmail.com), **Bushmeleva N. A.** (bushmeleva_na@list.ru), **Razova E. V.** (razova.ev@gmail.com), **Peskisheva T. A.** (peskisheva.t@mail.ru), **Pletneva M. V.** (pletneva.mv.kirov@gmail.com), Vyatka State University, Kirov, Russia

The sentiment lexicons are an important part of many sentiment analysis systems. There are many automatic ways to build such lexicons, but often they are too large and contain errors.

The paper presents the algorithm of sentiment lexicons creation for a given domain based on hybrid—manual and corpus-based—approach. This algorithm is used for the development of the sentiment lexicons by means of four human annotators each for five domains—user reviews of restaurants, cars, movies, books and digital cameras. Created sentiment lexicons are analyzed for inter-annotator agreement, parts of speech distribution and correlation with automatic lexicons.

The performance of the sentiment analysis based on the created sentiment lexicons is researched and compared with the performance of the existing sentiment lexicons. The experiments with text corpora on various domains based on SVM show high quality and compactness of the human-built lexicons.

THE PROBLEM OF SPEAKER IDENTIFICATION IN WHISPERED SPEECH

Kreychi S. A. (kreychi@mail.ru), **Krivnova O. F.** (okrivnova@mail.ru), **Stupina E. A.** (ek.stupina@gmail.com), Moscow State Lomonosov University, Moscow, Russia

The problem of speaker identification in whispered speech is of some interest for cognitive science, as well as for forensic and language minority. The work is devoted to an experimental study of the problem of familiar and unfamiliar speaker identification in vocal speech and in whisper. The experiment simulated in some respects the task and conditions of identification of a speaker by a listener. The initial number of participants—18 persons has been increased with the help of a questionnaire survey on the Internet. The number of remote auditors amounted to 125 people. The experiment took place in "online". The subject listened to and identified at least 16 pairs of entries. At the end of the experiment, he indicated if he had recognized any of the speakers. It was found that the main clues for correct recognition of a familiar speaker are the individual characteristics of his articulation, the components of the extraverbal part of his "speech portrait". The other features, such as individual speech style and individual manner of pauses and text macrosegmentation did not have any significant effects on speaker identification in whispered speech.

THE NATURAL LANGUAGE AND THE LANGUAGE OF GEOMETRIC SKETCHES

Kreydlin G. E. (gekr@iitp.ru), **Shabat G. B.** (george.shabat@gmail.com), The Russian State University for the Humanities, Moscow, Russia

The paper deals with the problems of interaction between the natural language together with its analogue—the natural-like language of geometry — and the language of geometric sketches within the two domains of intellectual activities. These are (1) synthesis and analysis of the natural language texts together with the corresponding non-verbal signs, and (2) oral and written multimodal communicative acts. Some linguistic (morphologic, syntactic and semantic) as well as semiotic peculiarities (the use of special signs, font markup, color, etc.) of the languages are discussed. The correspondence between some fragments of the natural-like language of geometry and some sketches is established. The problem of the representations of logical connectives and quantifiers in the sketches is partially solved by constructing the sketch analogs of the natural-like units and their combinations. It is stated that the more profound understanding of geometric facts and problems can be achieved by fluent knowledge of both languages and the special translation skills.

PROSODIC PHRASING IN SPOKEN TEXT: LOCALIZATION OF BREATHING PAUSES

Krivnova O. F. (okrivnova@mail.ru), Lomonosov Moscow State University, Moscow, Russia

In this paper we discuss the results of speech breathing research, undertaken to expand an empirical base for modeling of prosodic phrasing in Russian speech, The introductory section provides a brief description of the background, clarifies basic terms, explains the concept of breathing pause (BP) and its correlation with prosodic breaks and prosodic phrasing. In the second section we formulate the problems, discussed in this paper, with the main task to analyze the correlation of BP with the boundaries of the principal text units—paragraphs, sentences, clauses, taking into account the interspeaker variability in reading of the same text. The third section describes the material and methods of experimental analysis with particular attention to the possibilities of the computer detection of BP in a spoken text, as well as to the material adequate to the study. The fourth section outlines the general features of speech breathing in reading of the same text by different speakers. It is shown that one of the most common features is a different number of BPs that speakers make when reading the same text. It was also found that this variability is not related to the gender characteristics of the speakers or their place in the ranking of the best set of readings. Some correlation was found with the individual speech rate—the number of syllables spoken per second. However, despite this variability, all speakers use intonation pauses in the experimental text for breathing rather often. BP part of total intonational pauses averages 62% in the range from 52% to 74% by different speakers. The specific use of BP consists in the fact that they reflect the hierarchical structure of the text, with the individual clauses as the basis of it. Namely, text units, the end of which is accompanied by BP, are arranged in the direction of decreasing the probability of BP as follows (in parentheses the frequency of BP averaged by 10 speakers is given): paragraph (100%) > sentence inside a paragraph (94%) > clause inside a sentence (65%) > component in a clause (34%). In conclusion the study is summed up with the implication that BP in prosodic phrasing can serve as a sufficient signal of semantic text boundaries, but interspeaker variability shows that BP is not a necessary indication of them. The differentiating function of this prosodic marker is supported by the fact that BP with different text localization have stable differences in the overall phonetic picture and in such acoustic features as duration and intensity of breathing noise.

DISTRIBUTIVE BIPRONOMINAL CONSTRUCTIONS

Kustova G. I. (galinak03@gmail.com), Vinogradov Russian Language Institute of the Russian Academy of Sciences; Moscow State Pedagogical University (NSPU), Moscow, Russia

Pronominal complexes may be units of the pronouns' system (*drug druga*; *chto ugodno*; *neizvestno gde*) or may be constructions. Construction occupies an intermediate position between the free combinations and idioms. This paper discusses the bipronominal complexes (cf.: *Razbezhalis' kto kuda*), which have distributive meaning. Their first component is a distributive, the second one may be (1) interrogative pronoun (*Kto **kogda** priehal?* — 'When everyone has arrived'); (2) indefinite pronoun (*Zanimayutsya kto **chem*** 'everyone does something'), (3) relative pronoun (*Pomogaet komu **chem** mozhet* 'He helps everyone with what may').

Unlike other constructions with distributive semantics (cf.: *Chemodany popadali na pol* 'Suitcases fell to the floor'; *Kazhdy poluchil po 10 tsentov* 'They received ten cents each'), distributive bipronominal complexes are insufficiently investigated. The paper discusses the semantic and syntactic properties of bipronominal complexes — (2) and (3) types: are pronouns referential or non-referential; what determines the choice of pronouns' cases.

LEXICALIZED PROSODY AND THE POLYSEMY OF DISCOURSE MARKERS

Levontina I. B. (irina.levontina@mail.ru), Russian Language (Vinogradov) Institute, Moscow, Russia

The paper is devoted to some polysemous Russian particles and the problem of lexicalized prosody. The phenomenon of lexicalized prosody in Russian drew the linguists' attention about 30 years ago. The investigation is usually confined to phrasal stress as the most frequently lexicalized and therefore the most lexicographically interesting prosodic pattern. However, as far as discourse particles are concerned, not only phrasal stress but also intonation pattern is of greatest interest. Two Russian discourse particles will be discussed. One of them is the particle *-to*. Its different usages imply very different prosodic patterns. The other one is *vot* which can be used not only as a demonstrative particle, but also as a xenomarker (quotation marker), which requires a specific prosody.

COMPARISON OF MELODIC PORTRAITS OF ENGLISH AND RUSSIAN DIALOGIC PHRASES

Lobanov B. M. (Lobanov@newman.bas-net.by), United Institute of Informatics Problems NAS Belarus, Minsk

This study is an extension of the author's works, presented at the "Dialogue 2014 and Dialogue 2015" conferences. According to the concept of universal melodic portrait (UMP), a phrase intonation can be described as a sequence of UMPs of accentual units (AUs) that make up the phrase. The present paper describes the results of pilot studies where melodic portraits for English and Russian language phrases were compared. The examined phrases were derived from simple situational dialogues and were spoken by native English and Russian speakers. The study was restricted only to phrases with a one-accent unit structure representing the three main types of phrase intonations: affirmative statements, special questions and general questions.

The described UMP model allows to investigate tonal differences within languages by applying precise quantitative assessments. The method can be used effectively for solving problems of language interference. Moreover, the UMP model could potentially find an effective application in foreign language studies. Using the appropriate software that realizes the described stages of UMP construction, a learner could be able to visually compare an intonation of the pronounced phrase with its target intonation portrait and work to eliminate a foreign accent by proper training.

WORD SENSE DISAMBIGUATION FOR RUSSIAN VERBS USING SEMANTIC VECTORS AND DICTIONARY ENTRIES

Lopukhin K. A. (kostia.lopuhin@gmail.com), Scrapinghub, Moscow, Russia;
Lopukhina A. A. (nastya-merk@yandex.ru), V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

Word sense disambiguation (WSD) methods are useful for many NLP tasks that require semantic interpretation of input. Furthermore, such methods can help estimate word sense frequencies in different corpora, which is important for lexicographic studies and language learning resources. Although previous research on Russian polysemous verbs disambiguation established some important and interesting results, it was mostly focused on reducing ambiguity or determining the most frequent sense, but not on evaluating WSD accuracy. To the best of our knowledge, there is no comprehensively evaluated method that can perform semi-supervised word sense disambiguation for Russian verbs. In this paper we present a WSD method for verbs that is able to reach an average disambiguation accuracy of 75% using only available linguistic resources: examples and collocations from the Active Dictionary of Russian and large unlabeled corpora. We evaluate the method on contexts sampled from the web-based corpus RuTenTen11 for 10 verbs with 100 contexts for each verb. We compare different variations of the method and analyze its limitations. Method's implementation and labeled contexts are available online.

CREATING RUSSIAN WORLDNET BY CONVERSION

Loukachevitch N. V. (louk_nat@mail.ru)[1], **Lashevich G.** (design.berg@mail.com)[2], **Gerasimova A. A.** (anastasiagerasimova432@gmail.com)[1], **Ivanov V. V.** (nomemm@gmail.com)[2], **Dobrov B. V.** (dobrov_bv@mail.ru)[1]
[1]Lomonosov Moscow State University, Moscow, Russia
[2]Kazan Federal University, Kazan, Russia

In this paper we have described the semi-automatic process of transforming the Russian language thesaurus RuThes (in version, RuThes-lite 2.0) to WordNet-like thesaurus, called RuWordNet. In this procedure we attempted to achieve two main characteristic features of wordnet-like resources: division of data into part-of-speech-oriented structures with cross-references between them and providing a set of relations similar to WordNet-like resources. The published version of RuWordNet contains more than 115 thousand Russian words and phrases presented in form of three lexical nets for nouns, verbs and adjectives. Between synsets such relations as hyponym-hypernym, meronymy, part-of-speech synonymy, antonymy are established. In the paper we compare web-page representations of RuThes 2.0 and RuWordNet. It can be seen that RuThes looks as an ontology describing concepts and their relations and RuWordNet looks as a net of words. Researchers can obtain both types of thesauri and compare them in applications. In future, we will continue to add new types of relations to RuWordNet including the domain relation, the cause relation, the entailment relation, etc.

SENTIRUEVAL-2016: OVERCOMING TIME GAP AND DATA SPARSITY IN TWEET SENTIMENT ANALYSIS

Loukachevitch N. V. (louk_nat@mail.ru), Lomonosov Moscow State University, Moscow, Russia; **Rubtsova Y. V.** (yu.rubtsova@gmail.com), A. P. Ershov Institute of Informatics Systems, Novosibirsk, Russia

In this paper we present the Russian sentiment analysis evaluation SentiRuEval-2016 devoted to reputation monitoring of banks and telecom companies in Twitter. We describe the task, data, the procedure of data preparation, and participants' results. At the previous evaluation SentiRuEval-2015, it was noticed that the presented machine-learning approaches significantly depended on the training collection, which was not enough for qualitative classification of the test collection because of data sparsity and time gap. The current results of the participants at SentiRuEval-2016 showed that they have made successful steps to overcome the above-mentioned problems by combining machine-learning approaches and additional manual and automatically generated lexical resources.

LEXICAL RESEARCH IN RUSSIAN: ARE MODERN CORPORA FLEXIBLE ENOUGH?

Lukashevich N. Y. (natalukashevich@mail.ru), Moscow State University, Moscow, Russia; **Klyshinsky E. S.** (klyshinsky@mail.ru), Keldysh IAM RAS, Moscow, Russia; **Kobozeva I. M.** (kobozeva@list.ru), Moscow State University, Moscow, Russia

The article discusses what modern tools offer for a corpus-based lexical research in Russian. As an example we analyzed how the adjective *gordy* 'proud' is used in modern news texts. We studied data from such resources as two general Russian language corpora (RNC, GICR) and a corpus of syntactic co-occurrences containing information on syntactic relations of words for Russian (CoSyCo). If a corpus includes a variety of genres and allows to make fine-grained distinctions between text sources, it helps to highlight important style- and genre-dependent differences. Our comparison has demonstrated that there are quite significant differences in the usage of *gordy* which become clear when we study general news and IT news corpora separately, however, in general they show certain similar tendencies. It is also shown that when more varied genres are taken into account it may make more visible such style and genre features which it is not so easy to notice otherwise.

WELCOME TO THE CLUB: DESIGNING THE INVENTORY OF SEMANTIC ROLES FOR ADJECTIVES

Lyashevskaya O. N. (olesar@yandex.ru), National Research University Higher School of Economics, Vinogradov Institute of the Russian Language RAS, Moscow, Russia; **Kashkin E. V.** (egorkashkin@rambler.ru), Vinogradov Institute of the Russian Language RAS, Moscow, Russia

The argument constructions of adjectives has largely been out of the scope of research on semantic roles both in theoretical and IT fields. Before adding the roles of adjectival arguments to the network of semantic roles it is important to determine whether the adjectival roles form a separate list or whether they can be seen as an extension of roles assigned to the patterns of verbs and nominalizations. We discuss the general principles of how the inventory of adjectival roles should be organized in comparison with the existing inventories of verbal roles. In order to verify our statements, we carry out an experimental survey aimed at measuring the similarity between adjectival and verbal roles. The results have shown that both semantic interpretation of roles and their typical morpho-syntactic expression are significant for the evaluation and should be taken into account in working out the inventory. Besides, the specificity of adjectives lies in their prototypical stative semantics, which favors some differences in assigning a semantic role as compared to verbs. The results of the survey also provide some evidence for verification and development the inventory of verbal semantic roles.

FORMAL MODELING OF CASE VARIATION: A PARAMETRIC APPROACH

Lyutikova E. A. (lyutikova2008@gmail.com), MSU/MSPU, Moscow, Russia

This paper aims at evaluating formal theories of case assignment with respect to their applicability to modeling of case variation. Crosslinguistically, differential case marking exhibit significant variation in many parameters, including licensing factors of case variation, correlation of case with linear position, and feeding of predicative or possessive agreement. In this paper, I consider the two most elaborated formal theories of case—the minimalist syntactic case theory

and the configurational case theory—and explore their expressive power in modeling various types of differential case marking. I show that none of the theories is superior to the other—rather, each of them naturally accommodates a specific type of case variation but is unsuitable to express the other types. The minimalist syntactic case theory is more flexible in that it is compatible with additional mechanisms deriving the morphologically observable case variation, and more restrictive in that it predicts the one-to-one correspondence between case assignment and agreement. The prime advantage of the configurational theory is that it can represent directly the non-local dependencies between case-marking of different arguments.

GRAMMATICAL DICTIONARY GENERATION USING MACHINE LEARNING METHODS

Mazurova M. (sleepofnodreaming12@gmail.com), Ashmanov & Partners, Moscow, Russia

For the last decade, grammatical dictionaries have become not only a thing of theoretical value but an essential tool used in many fields of applied linguistics. However, the procedure of manual creation of a grammatical dictionary remains time- and labor-consuming. In this paper, the two-stage algorithm of automatic dictionary compilation, not requiring annotated texts, is proposed. As the source data, this system requires a formalized grammar description and a frequency distribution of a relatively large (hundred thousand tokens) corpus. Extending the principles commonly applicable to Indo-European languages, the research focuses on machine learning methods of corpora-based dictionary formation. Four machine learning models—SVM, random forest, linear regression and perceptron—are tested on the material of four languages: Albanian, Udmurt, Katharevousa, and Kazakh, and compared to a heuristic approach. While the linear models proved to be ineffective, other models' results were more promising: in an experiment with training and test sets formed from the same language's material, random forest reached 63% F-score, and SVM's results were also overdoing the baseline, however, the random forest model was unsuccessful. The best classifier in case of training and test sets based on the material of different languages was SVM. As a by-product of the experiments, the restrictions on the input were postulated: the approach 'as is' is not applicable to languages where inflections are strongly homonymic, and, on the contrary, is promising applied to an agglutinative language.

POSSESSIVES IN PARALLEL ENGLISH-CZECH-RUSSIAN TEXTS

Nedoluzhko A. (nedoluzko@ufal.mff.cuni.cz), Charles University in Prague, Prague, Czech Republic; **Schwarz (Khoroshkina) A.** (annakhor@gmail.com), ABBYY; Moscow State University, Moscow, Russia; **Novák M.** (mnovak@ufal.mff.cuni.cz), Charles University in Prague, Prague, Czech Republic

We present a corpus-based analysis of the use of possessive and reflexive possessive pronouns in a newly created English-Czech-Russian parallel corpus (PCEDT-R). Automatic word-alignment was applied to the texts, which were subsequently manually corrected. In the word-aligned data, we have manually annotated all correspondences of possessive and possessive reflexive pronouns from the perspective of each analysed language. The collected statistics and the analysis of the annotated data allowed us to formulate assumptions about language differences. Our data confirm the relative frequency of possessive pronouns in English as compared to Czech and Russian, and we explain it by the category of definiteness in English. To confirm some of our hypotheses, we used other corpora and questionnaires. We compared the translated texts in Czech and Russian from our corpus to the original texts from other corpora, in order to find out to what degree the translation factor might influence the frequency of possessives.

RUSSIAN MINORITY LANGUAGES ON THE WEB: DESCRIPTIVE STATISTICS

Orekhov B. (nevmenandr@gmail.com), **Krylova I.** (krylova93@gmail.com), **Popov I.** (imvanya@gmail.com), **Stepanova E.** (stepanovayekaterina@gmail.com), **Zaydelman L.** (luda.zaidelman@yandex.ru), National Research University Higher School of Economics, Moscow, Russia

The paper presents quantitative data about the web segments in minority languages of Russia. An ad-hoc search procedure allows to locate sites and pages on social networks that contain texts in a certain language of Russia. According to our data, there are texts in at least 48 of the examined languages on the Internet. We compared the gathered statistical data with the data from Wikipedia and the number of native speakers and found out that none of the "live" online data has a good correlation with the offline-life of language.

TOWARDS SEMANTICS OF THE RUSSIAN ASPECT: MOMENT OF PERCEPTION AND DISCURSIVE CONTEXT

Paducheva E. V. (elena.paducheva@yandex.ru), Federal Research Center Informatics and Management, Moscow, Russia

In the **context of a sentence** grammatical aspect (apart from its function of expressing multiplicity y) characterizes a situation with respect to the **moment of perception**. In the **context of discourse** the moment of perception also honestly does its job in the case of a sequence of identical aspectual forms. In fact, the notion "moment of perception" makes it possible to derive the textual (discursive) meaning of the Perfective and Imperfective forms from their meaning in an isolated utterance. The question arises: what happens in the context of juxtaposed or conjoined different aspectual forms. The subject of attention in the paper is a morphosyntactic configuration "Perfective verb + conjunction *i* 'and' + Imperfective verb". It is demonstrated that temporal relationships between situations in the context of this configuration heavily depend upon the moment of perception. If the perception moment of the Imperfective verb is synchronous to the perfective state of the Perfective one the **sequential** relationship between situations arises. If the perception moment of the Imperfective verb is synchronous to the event denoted by the Perfective verb the relationship between situations is **synchronicity**.

A DATABASE FOR VARIATIONAL STUDIES OF PALATALIZED/VELARIZED CONSONANTS PRECEDING [E] IN LOANWORDS

Perova D. M. (dmperova@mail.ru), **Bondarenko K. E.** (moidom@mail.ru), **Dobrushina N. R.** (nina.dobrushina@gmail.com), HSE, Moscow, Russia

The paper presents the initial/preparatory stage of the study of variation of hard/soft consonants before e in loanwords (*ka[f]e*). The main goal is to compile a database of relevant words for use in sociolinguistic research. The database is based on the list of word forms containing relevant contexts in users' queries to Yandex. All entries in the database are annotated for parameters that may be important in a variational study of the phenomenon. The article describes how the list was compiled and the principles of its annotation. The latter includes the consonant, the position of the consonant re the stressed syllable, the type of syllable where it occurs (open/closed), the year of the first occurrence of the word in Russian National Corpus; the language from which it was borrowed; its frequency. The database may be used to select stimuli for experimental studies of variation in modern speech and of its social correlates (age, gender, education, etc).

INTRA-SPEAKER STRESS VARIATION IN RUSSIAN: A CORPUS-DRIVEN STUDY OF RUSSIAN POETRY

Piperski A. Ch. (apiperski@gmail.com), Russian State University for the Humanities; National Research University Higher School of Economics, Moscow, Russia; **Kukhto A. V.** (anton.kukhto@gmail.com), Lomonosov Moscow State University, Moscow, Russia

Russian lexical stress exhibits both inter-speaker variation, defined by the speaker's regional affiliation, social status, age, etc., as well as intra-speaker variation. The latter is difficult to capture due to the need for large corpora of spoken text produced by one speaker. These are lacking, but can be replaced with poetic corpora. We use automatic analysis of poetic texts by 10 poets, drawn from the Russian National Corpus, in order to find word forms that can have stress variation. The number of such forms for an individual speaker ranges between 30 and 200 words, distributed among different parts of speech. We propose a quantitative measure of overall stress variability independent of the corpus size and show that there is a tendency for this variability to diminish over time, at least in poetic texts.

"NO PO RASCHOTU PO MOEMU DOLZHNA RODIT": THE RUSSIAN CONJUNCTION *NO* VIEWED THROUGH THE PRISM OF PROSODICALLY ANNOTATED CORPUS DATA

Podlesskaya V. I. (podlesskaya@ocrus.ru), Russian State University for the Humanities; Russian Academy of National Economy and Public Administration, Moscow, Russia

The paper focuses on Russian coordinate construction with clauses (or VPs) combined by means of the adversative conjunction NO. Prosodically, the construction may come up in two forms: (a) as a single illocution with the first clause pronounced with a rising pitch that projects discourse continuation, and (b) as two separate illocutions with the first clause pronounced with a falling pitch that projects no continuation. Basing on the data from the Prosodically Annotated Corpus of Spoken Russian, prosody and grammar of (a) and (b) were analyzed qualitatively and quantitatively. Type (b) appeared to be less frequent (comprising, approximately, 30% of the total number of occurrences) and systematically favored in contexts where the second clause is complicated by a "heavy" topical constituent.

VERBAL WORKING MEMORY AND SPEECH PRODUCTION DIFFICULTIES: DATA FROM RUSSIAN MULTIMODAL CORPUS

Potanina Y. D. (yulia.potanina.msu@gmail.com), Lomonosov Moscow State University, Moscow, Russia; **Podlesskaya V. I.** (vi_podlesskaya@il-rggu.ru), RSUH, RANEPA, Moscow, Russia; **Fedorova O. V.** (olga.fedorova@msu.ru), Lomonosov Moscow State University, Institute of Linguistics RAS, RANEPA, Moscow, Russia

It's a well-known fact that working memory capacity correlates with individual differences in comprehending speech (Daneman, Carpenter 1980). At the same time, the relationship between working memory capacity and speech production remains relatively unexplored. In this paper, we attempt to partially fill the gap and check the hypothesis about correlation between working memory capacity and number of lexical and grammatical markers of difficulties in production of spontaneous narratives. 19 Russian participants took part in two tests: the "Speaking Span" test by which we have measured their working memory capacity and the speech production test based on retelling the Pear Film (Chafe 1980). The Speaking Span test was designed in (Daneman and Green 1986) for English-speaking individuals: during the test increasingly longer sets of words are presented to participants; at the end of each set, they are supposed to use each word to generate a separate sentence (the word should be in the same grammatical form as it has been presented). Speaking span is measured as the maximum number of semantically and grammatically correct sentences produced in the experiment. This test was adapted to Russian: words in the set were balanced by syntactic categories, frequencies of individual lexemes and frequencies of grammatical forms. Collected narratives have been transcribed and manually annotated for lexical and grammatical markers of production difficulties. The documented number of lexical and grammatical markers of speech production difficulties varied between 0.77 and 8.58 per 100 words, which matches average rates reported previously in the literature. The study demonstrates the statistically significant correlation between working memory capacity measured by the "Speaking Span" test and verbal fluency measured in number of lexical and grammatical markers of production difficulties.

MACHINE-TRANSLATED TEXT DETECTION IN A COLLECTION OF RUSSIAN SCIENTIFIC PAPERS

Romanov A. V. (romanov@ap-team.ru), **Kuznetsova M. V.** (kuznetsova@ap-team.ru), **Bakhteev O. Yu.** (bahteev@ap-team.ru), **Khritankov A. S.** (khritankov@ap-team.ru), Antiplagiat.Research, Moscow, Russia; Moscow Institute of Physics and Technology (MIPT), Moscow, Russia

In this paper, we propose a method of machine-translated text detection. By 'machine-translated' texts, we mean, principally, the output of statistical machine translation systems. We focus on syntactic correctness and semantic consistency of sentences that constitute a text. More specifically, we make an attempt of detecting a certain phenomenon typically occurring in machine-translated documents. This phenomenon comprises the cases when small parts of the sentence, correctly translated, are combined together in an improper way. The proposed method is based on a supervised approach with a number of handcrafted features. First, we construct N-gram language models on a set of authentic scientific papers and on a set of machine-generated texts and assess the probability of each sentence according to these models. In

addition, we propose N-gram language models on part-of-speech tag sequences corresponding to the texts given. Furthermore, we explore the effectiveness of features obtained from two trained word2vec (CBOW and skip-gram) models. We assess quality of the method on a sample of Russian scientific papers, and English scientific documents machine-translated into Russian. Preliminary results demonstrate feasibility of the approach.

AUTOMATIC MORPHOLOGICAL TAGGING OF RUSSIAN SOCIAL MEDIA CORPORA: TRAINING AND TESTING

Selegey D. (danila-slg@yandex.ru)[1], **Shavrina T.** (rybolos@gmail.com)[1],
Selegey V. (Vladimir_S@abbyy.com)[2,3], **Sharoff S.** (s.sharoff@leeds.ac.uk)[4]
[1]Moscow State University, Russia; [2]Russian State University for the Humanities, Russia;
[3]ABBYY, Russia; [4]University of Leeds, UK

This paper presents a new set of basic tools for morphosyntactic tagging of Russian texts coming from social media. This has been developed within GICR, a project for creating a very large corpus of the Russian-speaking Internet.

This toolset includes a new tagset, obtained via extending and adaptating the tagset proposed by Sharoff et al. It has been tested on a gold standard test corpus of modern social media of about 2 million tokens. A particular feature of our approach is a fully automated process for development of training corpora. Instead of manual annotation we started with the output of the syntactic parser of Compreno. This annotation has been subsequently improved by automatic correction of systematic errors detected through processing of texts from social media. In this paper we show that existing tagging tools (in particular, tnt) produce consistently better results if they are trained with our corpus rather with other available corpora, in particular, those using the disambiguated portion of the Russian National Corpus.

The resulting test corpus is available in open access.

DISCURSIVE WORDS AND COMMUNICATIVES

Sharonov I. A. (Igor_sharonov.@mail.ru), RSUH, Moscow, Russia

The article is devoted to communicatives — special use of words, idioms and short sentences in dialogical positions of stereotype responses (response particles), intended to agree, disagree, answer some etiquette formulae, or to express different emotions. These are conversational formulae like *Da; Net uzh; Kakoe tam; Obaldet'!; Na zdorovje!*, etc. Communicatives consist of particles and idiomatic constructions; they are semantically empty and pragmatically specific. These units are regularly used in conversation, but not a single Russian dictionary has yet seen light where one could get compete information about communicatives and their occurrence in conversation. Only very few communicatives can be found in explanatory dictionaries and dictionaries of idioms. But their description in those rare cases is limited by their intention (affirmative response, doubt, etc.) or their function — either an etiquette answer (used to express thanks, regret) or an emotional response (used to express surprise, joy, grief). In some cases, communicatives may be marked as synonyms. Some words and idioms may function both as discourse words and communicatives, and some modern dictionaries claim to contain full information concerning their semantics and use. But the attention in the dictionaries is focused mostly on narrative, not dialogical, contexts, which distorts communicatives' actual use. The objective of the article is to compare characteristics of discursive words and communicatives. In the few examples we try to demonstrate the differences between the meaning and usage of these units and argue for the compiling the special Dictionary of Communicatives.

THE MOST FREQUENT WORDS IN EVERYDAY SPOKEN RUSSIAN (IN THE GENDER DIMENTION AND DEPENDING ON COMMUNICATION SETTINGS)

Sherstinova T. Yu. (t.sherstinova@spbu.ru), Saint-Petersburg State University,
Saint-Petersburg, Russia

The paper presents the most frequent words of everyday spoken Russian, that form the upper zones of several word frequency lists compiled on the material of Russian speech corpus "One Speaker's Day" (the ORD corpus), containing real-life recordings of everyday communication. All speech data in the corpus is annotated in terms of communication settings, including 1) type

of communication (language spoken style), 2) social role of speaker, 3) locus, etc. Such information allows speech to be filtered upon user request and therefore makes it possible to study speech variation depending on particular communication settings. The given study was made on the transcripts of 152 real-life macroepisodes and contains 232,370 words. The sample presents speech of 209 persons (95 men, 94 women, 20 children). The following word frequency lists have been compiled: a) general frequency list, b) male frequency list, c) female frequency list, and d) four frequency lists for different styles of spoken speech: informal conversations, professional/business conversations, educational communication, and "customer-service" communication. Men's and women's frequency lists have been compiled on the subsamples of 83,371 and 115,110 words correspondently. The analysis of word lists has shown that Russian women pay more attention to maintaining the conversation, use fewer hesitations, and are more inclined to use in their speech intensifying words, emotional words, hedges and interjections. Men generally use fewer personal pronouns, while numbers and the expletives are among the most frequent words used by men in everyday conversations. In general, these observations are similar to those described earlier for gender variation by other linguists.

MULTI-PRONUNCIATION LEXICON FOR RUSSIAN AUTOMATIC SPEECH RECOGNITION (PILOT STUDY)

Shirokova A. (anna_a@stel.ru), **Telesnin B.** (telesnin_ba@stel.ru),
Rogozhina V. (mind_your_own_business@rambler.ru), Stel CS, MSLU, Moscow, Russia

Our pilot study is aimed at building a lexicon of effective pronunciation variants on the basis of canonical pronunciations, for implementing it into the automatic speech recognition system for Russian. We focus on phonetic changes in word pronunciation caused by different factors operating in spontaneous speech. Our speech data includes three different corpora of the conversational type. Manual expert processing and analysis of the audio data are used. The lexicon construction procedure is given. Some statistics for pronunciation variation in Russian, obtained from the speech data, is presented. A description of frequent types of this phenomenon is given. Parallel and sequential pronunciation variants are discussed. Ways of formulating general phonetic variation rules and predicting potential contexts, in which pronunciation variation is likely to appear, are considered. Test data, phoneset used, and automatic speech recognition (ASR) parameters are described. Preliminary results for ASR and key word spotting (KWS) are shown. The appropriateness of using multi-pronunciation lexicon is discussed.

BELARUS VS. *BELORUSSIA*: THE STRUCTURE OF A LINGUO-POLITICAL CONFLICT IN SOCIAL MEDIA

Somin A. A. (somin@tut.by), National Research University Higher School of Economics, Moscow, Russia; **Poliy A. A.** (a.a.polyi@gmail.com), Russian State University for the Humanities, Moscow, Russia

This paper studies different aspects of a linguo-political conflict concerned with choosing between two Russian toponymic variants—*Belorussia* and *Belarus'* as well as adjectives *belorusskij (Belorussian)* and *belarus(s)kij (Belarusian)* and ethnonyms *belorus* and *belarus*. The core of the problem is that in the Russian language of Russia the variant *Belorussia* is used, which is considered to be insulting by many Belarusians, who prefer to use the variant *Belarus* while speaking Russian. In an attempt to understand the structure of this conflict, we analyze how and why the toponym *Belarus* appeared and spread through the newspapers of 1990-s, study the data from two online polls and the distribution of some words derived from the two toponymic variants, and finally discuss the scenarios of conflict communication in discussions in various social media. One of the polls shows the social distribution of the two toponymic variants and the other examines the attitude of the Belarusians towards the toponym *Belorussia* and its derivates. We show that each side of the conflict has its own limited set of ideas that reappear in conflict communication in comments under different articles on the Internet.

SPELLRUEVAL: THE FIRST COMPETITION ON AUTOMATIC SPELLING CORRECTION FOR RUSSIAN

Sorokin A. A. (alexey.sorokin@list.ru)[1,3,4], **Baytin A. V.** (baytin@yandex-team.ru)[2], **Galinskaya I. E.** (galinskaya@yandex-team.ru)[2], **Shavrina T. O.** (rybolos@gmail.com)[1,4]
[1]Lomonosov Moscow State University, Moscow, Russia; [2]Yandex, Moscow, Russia; [3]Moscow Institute of Physics and Technology, Dolgoprudny, Russia; [4]General Internet Corpora of Russian, Moscow, Russia

This paper reports on the first competition on automatic spelling correction for Russian language—SpellRuEval—held within the framework of "Dialogue Evaluation". The competition aims to bring together groups of Russian academic researchers and IT-companies in order to gain and exchange the experience in automatic spelling correction, especially concentrating on social media texts. The data for the competition was taken from Russian segment of Live Journal.

7 teams took part in the competition, the best results were achieved by the model using edit distance and phonetic similarity for candidate search and n-gram language model for their reranking. We discuss in details the algorithms used by the teams, as well as the methodology of evaluation for automatic spelling correction.

AUTOMATIC DETECTION OF MORPHOLOGICAL PARADIGMS USING CORPORA INFORMATION

Sorokin A. A. (alexey.sorokin@list.ru)[1,2], **Khomchenkova I. A.** (irina.khomchenkova@yandex.ru)[1]
[1]Lomonosov Moscow State University, Moscow, Russia; [2]Moscow Institute of Physics and Technology, Dolgopudnyj, Russia

This paper deals with automatic induction and prediction of morphological paradigms for Russian. We apply a method of longest common subsequence to extract abstract paradigms from inflectional tables. Then we experiment with the automatic detection of paradigms using a linear classifier with lexeme suffixes and prefixes as features. We show that Russian noun paradigms could be automatically detected with 77% accuracy per paradigm and 93% accuracy per word form, for Russian verbs per-paradigm accuracy reaches 76% and per-form accuracy is 89%. Usage of corpora information and character n-grams allows to improve these results up to 82% and 95% for nouns and 86% and 95% for verbs.

AUTOMATIC SPELLING CORRECTION FOR RUSSIAN SOCIAL MEDIA TEXTS

Sorokin A. A. (alexey.sorokin@list.ru)[1,2,3], **Shavrina T. O.** (rybolos@gmail.com)[1,3]
[1]Lomonosov Moscow State University, Moscow, Russia; [2]Moscow Institute of Science and Technology, Dolgoprudny, Russia; [3]General Internet Corpus of Russian, Moscow, Russia

This paper describes an automatic spelling correction system for Russian. The system utilizes information from different levels, using edit distance for candidate search and a combination of weighted edit distance and language model for candidate hypotheses selection. The hypotheses are then reranked by logistic regression using edit distance score, language model score etc. as features. We also experimented with morphological and semantic features but did not get any advantage. Our system has won the first SpellRuEval competition for Russian spell checkers by all the metrics and achieved F1-Measure of 75%.

FACTRUEVAL 2016: EVALUATION OF NAMED ENTITY RECOGNITION AND FACT EXTRACTION SYSTEMS FOR RUSSIAN

Starostin A. S. (astarostin@abbyy.com)[1], **Bocharov V. V.** (bocharov@opencorpora.org)[2,3], **Alexeeva S. V.** (sv.bichineva@gmail.com)[2,3], **Bodrova A. A.** (anastasie.bodrova@gmail.com)[2,3], **Chuchunkov A. S.** (alex.chuchunkov@gmail.com)[2], **Dzhumaev S. S.** (sdzhumaev@abbyy.com)[1], **Efimenko I. V.** (veassi@mail.ru)[4], **Granovsky D. V.** (dima.granovsky@gmail.com)[2], **Khoroshevsky V. F.** (v.khor@mail.ru)[5], **Krylova I. V.** (krylova93@gmail.com)[2], **Nikolaeva M. A.** (mary.nikolaeva@gmail.com)[2], **Smurov I. M.** (ismurov@abbyy.com)[1], **Toldova S. Y.** (stoldova@hse.ru)[6]

[1]ABBYY, Moscow, Russia; [2]OpenCorpora.org; [3]St. Petersburg State University; [4]Semantic Hub, Moscow, Russia; [5]Dorodnicvn Computing Centre, RAS (CC RAS); [6]Russian Research University—Higher School of Economics

In this paper, we describe the rules and results of the FactRuEval information extraction competition held in 2016 as part of the Dialogue Evaluation initiative in the run-up to Dialogue 2016. The systems were to extract information from Russian texts and competed in two named entity extraction tracks and one fact extraction track. The paper describes the tasks set before the participants and presents the scores achieved by the contending systems. Additionally, we dwell upon the scoring methods employed for evaluating the results of all the three tracks and provide some preliminary analysis of the state of the art in Information Extraction for Russian texts. We also provide a detailed description of the composition and general organization of the annotated corpus created for the competition by volunteers using the OpenCorpora.org platform. The corpus is publicly available and is expected to evolve in the future.

INFORMATION EXTRACTION BASED ON DEEP SYNTACTIC-SEMANTIC ANALYSIS

Stepanova M. E. (Maria_Ste@abbyy.com)[1], **Budnikov E. A.** (Egor_B@abbyy.com)[1], **Chelombeeva A. N.** (Antonina_C@abbyy.com)[1], **Matavina P. V.** (Polina_Ma@abbyy.com)[1], **Skorinkin D. A.** (Daniil_S@abbyy.com)[1,2]
[1]ABBYY, Moscow, Russia; [2]Higher School of Economics, Moscow, Russia

This paper presents a rule-based approach to Information Extraction (IE) task within FactRuEval-2016 competition. Our system is based on ABBYY Compreno Technology. The technology uses the results of deep syntactic-semantic analysis, which leads to significant reduction of the number of necessary rules and makes them laconic.

The evaluation was conducted on FactRuEval dataset. FactRuEval is an open evaluation of IE systems. The participants could take part in three tracks. The first track required to detect the boundaries and type of named entities in a text. The second track required to extract normalized attributes and perform local identification of named entities. The third track required to extract facts of certain types from a text. We took part in all three of the tracks with the nickname *violet*. Our method proved to be successful: we have achieved high F-measures in Named Entity Recognition tracks and the highest F-measure in Fact Extraction track.

CONTROL IN PURPOSE CONSTRUCTIONS WITH CAUSATION-OF-MOTION VERBS IN RUSSIAN: EVIDENCE FROM RUSSIAN NATIONAL CORPUS

Stoynova N. M. (stoynova@yandex.ru), V. V. Vinogradov Russian Language Institute, Moscow, Russia

The paper deals with motion-cum-purpose constructions in Russian. The constructions "causation-of-motion verb + infinitive" (*vesti kogo-to delatj čto-to* — lit. 'to lead somebody to do something') are observed on the data of Russian National Corpus. The problem of infinitive control is in focus. The Ø-subject of infinitive can be coreferential with the subject or with the object of the finite verb in such a construction: cf. *on vedet jejo ubivatj* lit. 'he$_i$ is leading her$_j$ Ø$_i$ to kill' vs. *on vedet jejo umiratj* lit. 'he$_i$ is leading her$_j$ Ø$_j$ to die'. The frequency of uses with subject-control correlates with the degree of formal and semantic cohesion within the purpose construction. One of the parameters is word order (which reflects the degree of syntactic cohesion). The second one is the spatial type of causation-of-motion verb: neutral unprefixed (*vesti*) / lative prefixed (*privesti*) / elative prefixed (*otvesti*). It reflects somehow the degree of semantic cohesion (neutral > lative > elative). The choice of controller is conditioned also by the referential type of the object (person / non-person / inanimate) and by the semantics of purpose event.

NAMED ENTITY RECOGNITION IN RUSSIAN: THE POWER OF WIKI-BASED APPROACH

Sysoev A. A. (sysoev@ispras.ru), **Andrianov I. A.** (ivan.andrianov@ispras.ru), Institute for System Programming of RAS, Moscow, Russia

Named entity recognition and classification is an important natural language processing task, aimed at finding words and word sequences, which denote named entities of different types in plain texts. This challenge was addressed in Task 1 of FactRuEval-2016 evaluation.

In the context of this evaluation, our team, acting for the Institute for System Programming of the Russian Academy of Sciences, proposed two approaches to exploiting information, mined from Wikidata and Wikipedia, for improving quality of named entity detection methods. In the first approach word2vec word embeddings, computed on Wikipedia, are used along with basic features in tokens classification. The second approach utilizes both Wikipedia and Wikidata to automatically construct a representative corpus for named entity recognition and classification training. Additionally, Wikidata, treated as a property graph, is used to collect named entity specific word dictionaries.

Our approaches (marked with identifier 'Orange' in FactRuEval-2016 organizers' quality evaluation reports) show up promising results, doing especially good for such well-defined class as person, still being appropriate for detecting named entities of other types as well.

"APPOSITIONAL" AND "CO-DETERMINATIVE" CONDITIONAL CLAUSES: IN SEARCH OF THE LOCUS OF CONDITIONALITY

Tiskin D. B. (daniel.tiskin@gmail.com), St. Petersburg State University, St. Petersburg, Russia

The paper provides evidence for the claim that the Russian conditional conjunction *esli* 'if' is itself devoid of either conditional or determiner semantics. The argument proceeds as follows. I demonstrate that with conjoined conditionals, just like with some NPs/DPs and with free relatives, one gets not only the immediately obvious "parallel" reading ('for *A* and for *B*') but also the "co-determinative" reading ('for those *A* which are also *B*'). The sort of reading identified in the literature as "appositional" turns out to be a subclass of co-determinative readings.

It has been proposed that appositional readings for NPs/DPs result from the fact that the pertinent DPs denote properties, whereas their conversion into referring or quantificational expressions is performed by type-shifting rules. Applying the same technique to conditionals, I conclude that the conditional *esli* cannot have the semantics of the definite determiner in the domain of possible worlds. Given the influential view (Lewis etc.) that *if* does not quantify over worlds either (that work done rather by "adverbs of quantification", which may be overt, e. g. *always* or *usually*, or covert), *esli* ends up free from any semantic duty, except that—as I argue—it determines whether the quantification over worlds or over time instances takes place (cf. *esli* vs. *kogda* 'when').
The proposed analysis may be used as guidance for the development of automatic recognition and analysis rules for such constructions.

COREFERENCE IN RUSSIAN ORAL MOVIE RETELLINGS (THE EXPERIENCE OF COREFERENCE RELATIONS ANNOTATION IN "RUSSIAN CLIPS" CORPORA)

Toldova S. Yu. (toldova@yandex.ru), **Khudyakova M. V.** (mkhudyakova@hse.ru), **Bergelson M. B.** (mbergelson@hse.ru), National Research University "Higher School of Economics", Moscow, Russia

The work deals with the experiment on adapting the annotation for coreference in written texts in Russian (used in Russian coreference corpus RuCor) to the corpus of Russian oral narratives from the Russian Clinical Pear Stories Corpus (Russian CliPS) (Bergelson et al., 2015). Russian CLiPS is a corpus of Russian "Pear stories" movie (Chafe, 1980) retellings in clinical populations in comparison with neurologically healthy people. The analysis deals with 10 texts by healthy people and 5 texts by people with various types of aphasia. The focus is on the specificity of reference choice in oral retellings and the specific features that should be taken into account for the annotation procedure for this type of data. The specific features for annotation of referential choice in clinical populations are also under discussion. The main claims are as follows. Some types of speech disfluencies should be integrated into the coreference annotation scheme. These

are noun phrases, which are repetitions of a previous referent mention, a referent renaming, or name correction. On the one hand, such noun phrases can influence the referent activation; on the other hand, they could lit a light on the process of the referential expression choice. Some of the referential expressions in spontaneous speech have epistemic markers as modifiers. They are also should be taken into consideration for the same reason. Thirdly, the NP structure and zero-anaphora should have more granulated set of features for coreference devices in oral speech, which are more diverse as compared to written ones. Moreover, some types of structures, such as adjectives postposition etc. and some types of zeros are the specific features of referential expressions in spontaneous speech.

MULTIPLE FEATURES FOR MULTIWORD EXTRACTION: A LEARNING-TO-RANK APPROACH

Tutubalina E. V. (tutubalinaev@gmail.com), Kazan Federal University, Kazan, Russia;
Braslavski P. I. (pbras@yandex.ru), Ural Federal University, Yekaterinburg, Russia

This paper describes the extraction of multiword expressions (MWEs) from corpora for inclusion in a large online lexical resource for Russian. The novelty of the proposed approach is twofold: 1) we use two corpora—the Russian National Corpus and Russian Wikipedia—in parallel and 2) employ an extended set of features based on both data sources. To combine syntactic and statistical features derived from two corpora, we experiment with several learning-to-rank (LETOR) methods that have been proven to be highly effective in information retrieval (IR) scenarios. We make use of bigrams from existing dictionaries for learning, which leads to very sparing manual annotation efforts. Evaluation shows that machine-learned rankings with rich features significantly outperform traditional corpus-based association measures and their combinations. Analysis of resulting lists supports the claim that multiple features and diverse data sources improve the quality of extracted MWEs. The proposed method is language-independent.

ASPECTUAL PAIRS, SEMANTIC THEORY AND MASLOV CRITERION

Uryson E. V. (uryson@gmail.com), Russian Language Institute (Russian Academy of Sciences)

The paper deals with Russian aspectual pairs like umirat'—umeret', risovat'—narisovat' (but not obizhat'sa—obidet'sa). The imperfective verb in a pair designates a process or an action, while the perfective verb designates the "resulting event" completing the process / action. It is well known that in some diagnostic contexts the imperfective member of a pair substitutes it's perfective correlate and thus designates an event (Maslov criterion). It will be demonstrated that this substitution is due to certain semantic components in the lexical meaning of a verb. Fot this purpose the progressive meaning of imperfective verbs will be analysed. We will argue that the component 'the resulting event' is a part of meaning both of a perfective verb and the imperfective one in an aspectual pair. Status of this component in the lexical meaning of an imperfective verb will be discussed. Maslov's diagnostic contexts will be observed. A criterion for determing the imperfective correlate to a given perfective verb in some controversial cases (cf. est' / s'edat'—s'est') will be suggested in addition to Maslov criterion.

COMPARING CORPUS AND EXPERIMENTAL APPROACHES: A STUDY OF SYNTACTIC PROPERTIES OF THE RUSSIAN ENCLITIC *ŽE*

Valova E. A. (dunya_v@yahoo.com), HSE, Moscow, Russia; **Slioussar N. A.** (slioussar@gmail.com), HSE, Moscow; St. Petersburg State University, St. Petersburg, Russia

Corpus and experimental approaches in linguistics are often seen as incompatible, and there are very few studies of grammatical phenomena that rely on both of them, without one or the other being subsidiary. In this paper, we would like to show that they are complimentary and can be fruitfully combined on the example of Russian phrasal enclitic *že*. We analyze various factors influencing its position in the sentence, in particular, whether it obeys Wackernagel's law, which applied to phrasal enclitics in Old Russian.

Data from the National Russian corpus show that *že* appears in the strict Wackernagel's position in the absolute majority of cases in the main subcorpus and the newspaper subcorpus, while the subcorpus of spoken Russian exhibits more variation. Corpus data allow tracing

diachronic tendencies and identifying several factors (primarily, the semantics of *že*). Experimental data let us estimate the role of these and other factors on a carefully balanced set of examples. Apart from syntactic and semantic factors, the age and educational level of participants was demonstrated to influence the results.

"AS THEY SAY, AN ARTICLE IS AN ARTICLE": SOME ASPECTS OF USE OF TAUTOLOGIES IN COMMUNICATION

Vilinbakhova E. L. (e.vilinbakhova@spbu.ru), St. Petersburg State University, St Petersburg, Russia

In most studies dedicated to tautologies it goes without saying that these constructions are commonly used in everyday speech. Further analysis is based on the hearer's point of view concentrating mostly on possible ways of interpretation of tautologies. At the same time the perspective of the speaker remains largely unexplored. This study based on Internet and corpus data [RNC] deals with some aspects of use of tautologies in communication in order to understand why the speaker should opt for uttering tautologies instead of being more straightforward and what communicative profit he gets for that. It seems that advantages of using tautologies for the speaker are based on their structural and semantic features: (a) their recognizable form *X cop X* that makes tautologies look like a cliché; (b) their possibility to appeal to mutual knowledge; (c) the unquestionable truth of their literal meaning. First, when the speaker uses tautologies as clichés with expressions "as they say", *etc*., he makes his personal opinion look like a common wisdom of linguistic community. Next, when the speaker emphasizes that he appeals to mutual knowledge, he makes the hearer look as like-minded person, therefore the hearer's possible disagreement is regarded as a refusal of (expected) support and solidarity and requires more effort. Finally, the fact that the literal meaning of tautologies is undeniable helps the speaker escape of the responsibility of false implicature; defend his opinion using so-called *deep tautologies*; close the discussion whenever it is more convenient to him.

THE ROLE AND APPLICATIONS OF EXPERT ERROR ANNOTATION IN A CORPUS OF ENGLISH LEARNER TEXTS

Vinogradova O. I. (olgavinogr@gmail.com), Research University Higher School of Economics, Moscow, Russia

The paper presents the rationale for the decisions that were taken in the set-up and further development of a learner corpus of student texts written in English by Russian learners of English, the only Russian learner corpus in the open access. The tool of manual expert annotation is in the focus of the present observations, and after introducing categorization of errors applied in annotation, the complicated cases that arose in annotation practices have been looked into followed by comparison of the annotation statistics over the three stages in the corpus development. For that purpose, texts annotated by different groups of participants in the process of two experiments were used to spot the problematic areas in annotation. The main pedagogical applications of the learner corpus in teaching EFL—the opportunities to create automated training exercises and placement and progress tests custom-made for specific groups of students—are outlined in the concluding part of the paper.

PROSODIC TRANSCRIPTION FOR THE NEW RUSSIAN PROSODIC CONSTRUCTIONS

Yanko T. E. (tanya_yanko@list.ru), Institute of Linguistics, Moscow, Russia

This paper is aimed at investigating two Russian pitch accents never discussed in the Russian linguistics. These are: the gradual rise found in one the Russian regional variants, namely in Odessa Russian, and the prosody of breaking information into portions in standard literary spoken Russian. The gradual rise has a rising tone on the tonic syllable and gradually rising frequency changing on the post-tonic syllables. The prosody of breaking information into portions has a falling tonic syllable and slightly rising post-tonics. It also has a prolonged time of articulation. The tonal and temporal parameters of the pitch accents in consideration, their functions in discourse, and their phonological status are discussed. The criterion for the pitch accent to be

viewed as an autonomous phonological unit of a language is whether the pitch accent has permanent means of expression and a stable function, or a limited set of functions in discourse. For describing the newly introduced pitch accents, the transcription based on the Pierrehumbert's autosegmental notation of prosody was used. For investigation, a minor working corpus of the Russian speech recordings was set up. It comprises two components. The first component of the corpus consists of short stories about Odessa told by the citizens, jokes, and funny stories. The second component includes recordings of friendly talks and radio conversational programs in standard Russian. The software program Praat was used in the process of analyzing the sounding data. The results presented here are exemplified by frequency tracings of records taken from the corpus.

A DATABASE OF CROSS-LINGUISTIC EQUIVALENCES AS AN INSTRUMENT OF LINGUISTIC ANALYSIS

Zalizniak Anna A. (anna.zalizniak@gmail.com), Institut of linguistics RAS, Moscow; Institute of Informatics Problems FRC CSC RAS, Moscow

The paper outlines the principles of creation of a Database of Russian language-specific units and their French equivalents and the possibilities of its use as a tool of linguistic analysis. The entry of the Database is a monoequivalence (ME), i. e. a dyadic tuple, which consists of a Russian sentence including a language-specific unit and its French translation (automatically extracted from the Russian-French subcorpus of Russian National Corpus), including a functionally equivalent fragment (FEF) of the Russian language-specific unit. Both constituents of the ME are annotated with two-level characteristics, ensuring their faceted classification: "basic type" and "additional feature". The paper indicates relevant quantitative parameters that can be extracted from such a database and can be accounted for in the analysis of language-specific units; it demonstrates that quantitative methods can be effectively used only in combination with proper methods of semantic analysis. The reliability of statistical data will increase with the extension of the volume of the parallel corpus.

ON THE ASPECTUAL STATUS OF CONATIVE PAIRS IN RUSSIAN: WHY *SEARCH* CANNOT MEAN *FIND*?

Zalizniak Anna A. (anna.zalizniak@gmail.com), Institute of Linguistics, RAS, Moscow, Russia;
Mikaelian I. L. (irina-mikaelian@yandex.ru), The Pennsylvania State University, USA

The discussion of the so-called conative pairs in the Russian language, i. e. pairs including an imperfective and a perfective verb where the former expresses the idea of an attempt to reach a goal while the latter designates a successful fulfillment of that goal, has a long tradition.

However, the aspectual status of such pairs remains unclear. In particular, the aspectual status of the verbs *iskat'*$_{ipf}$ — *najti*$_{pf}$ (*to search — to find*) has not found consensus among scholars of the Russian language. The paper provides answers to the following questions: Why the verbs *iskat' — najti* do not function as an aspectual pair in Russian? Why however Russian speakers perceive this pair as an aspectual one? How the non-aspectual pair *iskat' — najti* is related to conative aspectual pairs *lovit' — pojmat' и reshat' — reshit'*? How the non-aspectual pair *iskat' — najti* is related to telic aspectual pairs *razyskivat' — razyskat'* and *otyskivat' — otyskat'*?

Авторский указатель

Author Index

Научное издание

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной международной конференции «Диалог»

Выпуск 15 (22). 2016

Ответственный за выпуск **А. А. Белкина**
Вёрстка **К. А. Климентовский**
Дизайн обложки **А. А. Светличная**

Подписано в печать 12.05.2016
Формат 152 × 235
Бумага офсетная
Тираж 350 экз. Заказ № 72

Издательский центр «Российский государственный гуманитарный университет»
125993, Москва, Миусская пл., д. 6
Тел.: +7 499 973 42 06

Отпечатано с готового оригинал-макета в типографии
ООО «Издательско-полиграфический центр Маска»
117246, Москва, Научный пр-д, д. 20, стр. 9