Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной Международной конференции «Диалог» (2014)

Выпуск 13

Computational Linguistics and Intellectual Technologies

Papers from the Annual International Conference “Dialogue” (2014)

Issue 13

УДК 80/81; 004
ББК 81.1
К63

Программный комитет конференции выражает
искреннюю благодарность Российскому фонду фундаментальных
исследований за финансовую поддержку,
грант № 14-07-20065-г

Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2014 г.). Вып. 13 (20). — М.: Изд-во РГГУ, 2014.

Сборник включает 64 доклада международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2014», 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

13-й выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит избранные материалы 20-й Международной конференции «Диалог». В результате работы более 50 рецензентов для публикации в ежегоднике было отобрано 64 доклада, охватывающих различные направления исследований в области компьютерного моделирования и анализа естественного языка, представленные на конференции:

- Формальные модели языка и их применение в компьютерной лингвистике
- Модели и методы семантического анализа текста
- Лингвистические онтологии
- Гибридные технологии компьютерного анализа текстов
- Теоретическая и компьютерная лексикография
- Типология и компьютерная лингвистика
- Компьютерные лингвистические ресурсы
- Корпусная лингвистика: создание, разметка, методики применения и оценка корпусов
- Интернет как лингвистический ресурс. Лингвистические технологии в Интернете
- Машинный перевод текста и речи
- Лингвистический анализ речи
- Модели общения. Коммуникация, диалог и речевой акт
- Извлечение знаний из текстов
- Компьютерный анализ документов: классификация, поиск, анализ тональности и т. д.

Сегодня в мире проводится немало конференций по компьютерной лингвистике. У «Диалога», ведущей конференции в этой области в России, есть две уникальные особенности: она ориентирована прежде всего на компьютерный анализа русского языка, и, как это следует из самого названия конференции, она нацелена на глубокое взаимодействие современной лингвистической теории и инженерной практики.

Каждый год Программный комитет предлагает участникам актуальную «интегрирующую» тему в качестве доминанты конференции. В этом году такой доминантой стали «Вычислительные модели семантики». Понимание критической важности использования семантических моделей в задачах компьютерной лингвистики растет год от года. Но сами применяемые модели существенно различаются: наряду с традиционными лингвистическими и формально-логическими подходами бурно развиваются дистрибуционные, операциональные, онтологически ориентированные методы. На конференции и в настоящем сборнике представлены работы исследователей, работающих в этих направлениях.

В последние годы важнейшей миссией «Диалога» стала разработка и апробирование методик верификации результатов лингвистических исследований и сравнительных оценок эффективности систем анализа текстов на русском языке. Целью этой работы является разработка единых для авторов и рецензентов «Диалога» принципов «evaluation»: доказательства эффективности и адекватности полученных результатов.

Такие доказательства возможно получить только в результате проведения серьезных тестов в соответствии с разработанными методиками. Такие тестирования ежегодно проводятся в рамках «Диалога». В этом году тестировались системы анализа местоименной анафоры и кореферентных связей в тексте. Как обычно, в сборнике опубликованы работы участников тестирования и итоговая статья его организаторов.

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 языке).

За год, прошедший после последней конференции, «Диалог» понес невосполнимую потерю: несправедливо рано ушел из жизни наш коллега Илья Сегалович, роль которого в «Диалоге» была очень значительна. Этому замечательному человеку посвящена памятная статья его коллег, завершающая сборник.

Несмотря на традиционную широту тематики докладов, отобранных рецензентами в этом году, 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.

Программный комитет «Диалога»
Редколлегия ежегодника «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
Богуславский Игорь Михайлович	Политехнический университет Мадрида
Буате Кристиан	Гренобльский университет
Гельбух Александр Феликсович	Национальный политехнический институт, Мехико
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН
Кобозева Ирина Михайловна	Филологический факультет МГУ
Козеренко Елена Борисовна	Институт проблем информатики РАН
Корбетт Гревил	University of Surrey, UK
Кронгауз Максим Анисимович	Институт Лингвистики РГГУ
Лукашевич Наталья Валентиновна	НИВЦ МГУ
Маккарти Диана	Lexical Computing Ltd., UK
Мельчук Игорь Александрович	Монреальский университет
Ниврe Йоаким	Уппсальский университет
Ниренбург Сергей	Университет Нью-Мексико
Осипов Геннадий Семёнович	Институт программных систем РАН
Попов Эдуард Викторович	РосНИИ информационной техники и САПР
Раскин Виктор	Purdue University, USA
Селегей Владимир Павлович	Компания ABBYY
Хови Эдуард	University of Southern California
Шаров Сергей Александрович	University of Leeds, UK

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	Kontur Labs; Уральский федеральный университет
Добров Борис Викторович	НИВЦ МГУ
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН
Кобозева Ирина Михайловна	Филологический факультет МГУ
Козеренко Елена Борисовна	Институт проблем информатики РАН
Лауфер Наталия Исаевна	ООО «проФан Продакшн»
Ляшевская Ольга Николаевна	Universitet i Tromsø, Norway
Сердюков Павел Викторович	Компания Yandex
Соколова Елена Григорьевна	РосНИИ искусственного интеллекта
Толдова Светлана Юрьевна	Филологический факультет МГУ
Шаров Сергей Александрович	University of Leeds, UK

Секретариат

Белкина Александра Андреевна, *секретарь оргкомитета*	Компания ABBYY
Атясова Анастасия Леонидовна, *координатор*	Компания ABBYY

Рецензенты

Азарова Ирина Владимировна
Апресян Валентина Юрьевна
Байтин Алексей Владимирович
Баранов Анатолий Николаевич
Беликов Владимир Иванович
Богданов Алексей Владимирович
Богданова Наталья Викторовна
Богуславский Игорь Михайлович
Борщев Владимир Борисович
Браславский Павел Исаакович
Горностай Татьяна Александровна
Гришина Елена Александровна
Губин Максим Вадимович
Даниэль Михаил Александрович
Добров Борис Викторович
Добровольский Дмитрий Олегович
Добрынин Владимир Юрьевич
Дружкин Константин Юрьевич
Захаров Виктор Павлович
Иомдин Борис Леонидович
Иомдин Леонид Лейбович
Кобозева Ирина Михайловна
Крейдлин Григорий Ефимович
Кронгауз Максим Анисимович
Лахути Делир Гасемович
Левонтина Ирина Борисовна
Лобанов Борис Мефодьевич
Лукашевич Наталья Валентиновна
Ляшевская Ольга Николаевна
Маккарти Диана
Падучева Елена Викторовна
Пазельская Анна Германовна
Паперно Денис Аронович
Пиперски Александр Чедович
Плунгян Владимир Александрович
Подлесская Вера Исааковна
Рахилина Екатерина Владимировна
Савельев Василий Евгеньевич
Селегей Владимир Павлович
Сердюков Павел Викторович
Смирнов Иван Валентинович
Сокирко Алексей Викторович
Соколова Елена Григорьевна
Старостин Анатолий Сергеевич
Тестелец Яков Георгиевич
Тихомиров Илья Александрович
Толдова Светлана Юрьевна
Урысон Елена Владимировна
Федорова Ольга Викторовна
Филиппова Екатерина Александровна
Хорошевский Владимир Федорович
Циммерлинг Антон Владимирович
Шаров Сергей Александрович
Юдина Мария Владимировна
Янко Татьяна Евгеньевна

Содержание[1]

Основная программа конференции

---

[1] Доклады упорядочены по фамилии первого автора в соответствии с порядком английского алфавита.

* Тексты, отмеченные *, печатаются по решению Редсовета.

Contents[1]

Basic Conference Program

[1] The papers are ordered by the surname of the first author in compliance with the English alphabet.
* Papers marked by * are printed by the decision of the Editorial board.

Editorial

Основная программа конференции

AUTOMATIC CREATION OF HUMAN-ORIENTED TRANSLATION DICTIONARIES

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

Yandex, Moscow, Russia

This paper addresses the issue of automatic acquisition of a human-oriented translation dictionary from a large parallel corpus. Automatically generated dictionary entries can enrich the output of a statistical machine translation system. We describe an automatic approach to the extraction of translation equivalents, and dictionary entry construction: grouping of synonymic translations, selection of illustrative context examples. The extraction of possible translations is based on statistical machine translation methods. The selection of lemmatized and linguistically motivated phrases is done with the help of morpho-syntactic analysis. In contrast to human-built dictionaries, an automatic dictionary usually contains a certain amount of noisy translations, as a consequence of systematic alignment mistakes and corpus imperfections. A noise reduction approach is proposed. We also provide the result of an evaluation experiment and the comparison of frequency distribution of words in the queries to the dictionary and the frequency distribution of words in plain text.

Keywords: parallel texts, bilingual dictionary extraction

1. Introduction

This paper describes an approach to the automatic construction of translation dictionaries. The approach is based on statistical machine translation (SMT) methods and can be applied to various language pairs. The dictionary entries are created automatically and contain translation variants grouped by meaning, reverse translations and context examples. For some languages, the dictionary entries can also include data prepared partly manually, e.g. transcriptions.

The automatic acquisition of translation equivalents from parallel texts has been extensively studied since the 1990s [7, 14]. The set of translation pairs is often referred to as bilingual lexicon. At the beginning, the automatically acquired lexicons served as internal resources for SMT [3], information retrieval (IR) [15], or computer-assisted lexicography [2, 4].

The growth of Internet and the current progress in search of web-based parallel documents [10, 12] makes it possible to automatically construct large-scale bilingual lexicons. Hence a new interesting possibility arises—to produce automatically acquired human-oriented translation dictionaries that have a practical application.

A machine translation system can output an automatically generated dictionary entry in response to all queries that are found in the dictionary. The percentage of short queries can be quite large, and the system benefits from showing several possible translations instead of a single result of machine translation (Fig. 1).

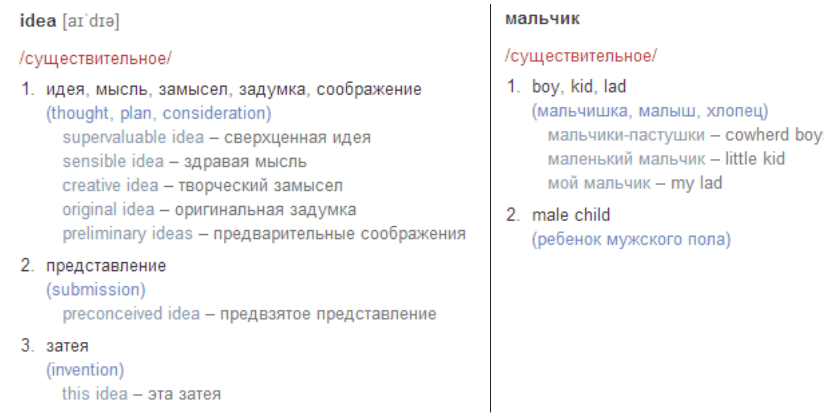


Fig. 1. Examples of dictionary entries in English—Russian and Russian—English dictionaries

The initial translation equivalents for an automatic dictionary bilingual lexicon can be extracted with the help of the techniques and tools developed for the phrase-table construction in SMT. The widely used word alignment and phrase extraction algorithms are described in [3, 9]. Though an SMT phrase-table actually consists of translation equivalents, it may differ substantially a human-oriented dictionary (Table 1). Additional algorithms are required to convert the initial translation equivalents into a dictionary.

Table 1. Differences between a human-oriented dictionary and an SMT phrase-table

Human-oriented dictionary	SMT phrase-table
Lemmatized entries are preferred.	Words and phrases in all forms are included.
Only linguistically motivated phrases are acceptable.	Any multiword combination is included.
Precision is important. Any noise is undesirable.	Having lots of low-probability noise is acceptable, since it is generally overridden by better translations.

The translation equivalents are organized into dictionary entries. The key of an entry is a word or phrase, usually lemmatized. Its translations are divided by the part of speech. Inside each part-of-speech class, the synonymic translations are grouped together. The groups are ordered according to their aggregate frequency.

Each group can be illustrated by reverse translations, and parallel context examples, drawn from the parallel corpus. Fig. 2 explains the structure of a dictionary entry for the word “French” in the English-Russian dictionary.

Fig. 2. The structure of a dictionary entry

Some aspects of the dictionary entry construction represent independent problems. The grouping of synonymic translations relies on the pre-constructed dictionary of synonyms, which also can be built automatically from the parallel corpus, as described in 2.5. The problem of selecting most illustrative context examples is discussed in 2.4.

In contrast to human-built dictionaries, an automatic dictionary usually contains a certain amount of noise—incomplete or totally incorrect translations. Yet, it may have some important advantages.

- Being objective and up-to-date. The frequency of uncommon or archaic translations is low. At the same time, the automatic approach often finds relevant translations, missed by a professional lexicographer [11].
- Improvement over time. With the possibility to process more parallel documents, the automatic dictionaries can potentially cover more words and phrases than the human-built dictionaries.
- Better flexibility. Since the procedure is fully automatic, it is easier to rearrange the dictionary, adjust its parameters (e.g. the precision/recall ratio, the maximum number of translations per a single entry). The translations can be ordered according to their frequencies or probabilities. This reduces the average time the user spends when looking for a particular meaning.
- Uniform approach to different language pairs.

The rest of the paper is organized as follows. We describe the overall system architecture in Section 2. We discuss the types of noisy translations and the noise detection approach in Section 3. The dictionary evaluation is described in Section 4. We conclude and discuss the applicability of the proposed approach to different language pairs in Section 5.

2. System Architecture

The overall process of the English-Russian dictionary construction is represented in Fig. 3.

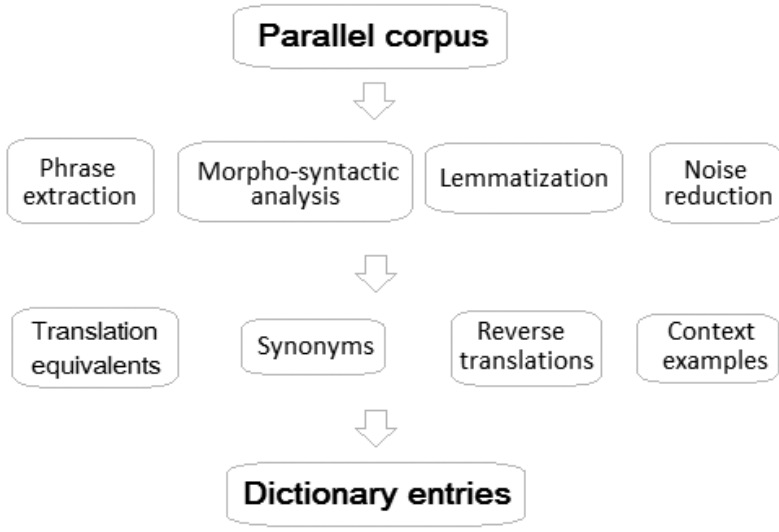


Fig. 3. The system architecture

2.1. Word Alignment, Morpho-Syntactic Analysis and Phrase Extraction

The parallel corpus is word-aligned and processed with English and Russian dependency parsers [1]. The initial phrase extraction is done as described in [8]. The maximum phrase length is limited by 3 words. We also discard the phrases where the words are not connected in any of the English and Russian parsing trees.

2.2. Lemmatization

The lemmatization is important; otherwise dictionary entries may contain many different forms of the same word, especially if the target language is morphologically

rich. The sentence preprocessing by an automatic lemmatization algorithm may introduce incorrect lemmas to the final dictionary. The reasons for that are ambiguity, heuristic lemmatization of unknown words, difficulties with the lemmatization of multi-word phrases. A possible way to overcome this problem is selecting a most lemma-like phrase pair among the real examples.

Each phrase pair can be assigned a key consisting of the lemmas of all words in it. The phrase pairs with the same key represent the translation equivalents in different forms. We select one best representative among them. The choice is made by taking into account the frequency of the unnormalized phrase pair and the morphological attributes.

2.3. Noise Reduction

Some undesired translation equivalents can be detected by simple heuristics. For example we remove the translations of punctuation, digits, or phrase pairs which occur extremely rarely. Still, there exist other types of noise which is more difficult to detect. We discuss this problem in detail in section 3.

2.4. Finding Context Examples and Reverse Translations to Illustrate the Meaning

Parallel context examples can help the user to distinguish the meaning of multiple translations. The examples are expected to be short well-formed grammatical phrases. We rely on the parsing trees to find such phrases. Besides, it is important to select not only the most frequent context example, but most distinctive and informative. This task is addressed by the metrics, such as mutual information, and the statistics of syntactic links between words.

Reverse translations also can be helpful to illustrate the word meaning.

2.5. Synonyms

The grouping of synonymic translations is done for two reasons. On the one hand, it visually shortens the list of translations and makes it easier to perceive. On the other hand, it also helps to differentiate faster between the different meanings.

The grouping procedure relies on the pre-constructed dictionary of synonyms, which is also built automatically from the same parallel corpus, from which the initial translation equivalents had been extracted. The idea behind automatic search of synonyms is that words with similar meaning are often translated by the same word in another language. We also used distributional similarity of syntactic contexts as an independent criterion of synonymity. Though each of the two factors—translation similarity and distributional similarity—could introduce incorrect synonyms, their intersection allows to increase the accuracy of the method.

2.6. Dictionary Construction for Different Language Pairs

The proposed procedure of dictionary construction requires morphological and parsing tools for both languages. Morphological tools are useful to determining possible parts of speech and lemmas of the word forms. Parsers are needed for the filtration of ungrammatical phrases and search of context examples. The part-of-speech disambiguation can be done within the parsing process, or with the help of a tagger. While the lack of such tools imposes certain restrictions on the dictionary content, the proposed approach can be still applicable in different cases.

If no parser is available, the dictionary will include neither multi-word translations, nor context examples. If morphology exists, but no morphological disambiguation tool is available, we can restrict the translations to those with identical part of speech.

3. Detection of Noisy Translations

The accuracy requirements are higher for a human-oriented dictionary, compared to a phrase-table used within an SMT system. The noise can appear as a consequence of systematic alignment mistakes and corpus imperfections, namely, nonparallel sentences, low-quality machine translation, language recognition mistakes. The following types of mistakes are common for automatically constructed dictionaries.

- Transliteration or translation by a word that belongs to a different language.
  челочка — chelochka

- Misspelled translation.
  monkey — обезъян

- Incomplete translation.
  доиграть — finish (finish playing)
  determined — определиться (be determined)
  present — памятный подарок (unforgettable present)

- Translation by an antonym. This can happen if one side incorporates a negation in its semantics, and the other sides has a negation as a separate word.
  eat — недоедать (be undernourished)
  unaware — подозревать (suspect)

- Translations of words with strict meaning. Though the highest-probability translations of proper names and colors are usually correct, some other variants may look unacceptable.
  yellow — белый (white)
  Russia — Украина (Ukraine)

- Translation by a common word.
  Russia — страна (country)

The most straightforward techniques of noise reduction in SMT phrase-tables is the filtration by frequency or probability thresholds [5]. However, in case of some systematic defects in the initial parallel corpus, a substantial amount of noise still survives, while many good translation equivalents are lost.

In addition to the translation probabilities, our approach to noise detection relies on the analysis of the parallel sentences in which a given translation pair occurred. There are several symptoms indicating that a sentence is a possible source of noisy translation:

- Unsafe one-to-one alignment. The intersection of HMM-based word alignments for two translation directions is a simple heuristic for finding the words that are confidently aligned to each other [9]. The percentage of such safe alignment points can vary in different sentences. However, its being too small is abnormal, and possibly indicates some defect to the given parallel sentences.
- High distortion of word order. Though some language pairs have different word order, the distance between the translations of subsequent input words is close to that of the input sentence. As well as unsafe one-to-one alignment, high distortion may indicate some defect to the given sentence pair.
- Bad syntactic structure. Some noisy translations originate from the sentences that seem to be an output of a poor-quality machine translation system. Bad translation often breaks up the syntactic structure of the output sentence.
- Many out-of-vocabulary words. Sentences containing many out-of-vocabulary words probably do not belong to the given language.
- Highly punctuated sentences. One can observe that sentences with lots of punctuation either are unnatural or contain enumeration. Large numeration lists are often not exactly parallel and can be aligned incorrectly, because many commas are mapped to each other.

4. Dictionary Evaluation

The evaluation of dictionary quality and the comparison of different dictionaries is a complicated task. Specifically, Tomaszczyk [13] considers multiple criteria for bilingual dictionary evaluation: equivalents, directionality, reversibility, alphabetization, retrievability, redundancy, coverage, currency, reliability. But these criteria are mostly qualitative and serve as a recommendation for a human expert reviewing a new dictionary.

One can also apply the standard information retrieval metrics, such as recall and precision. In this case, the manual gold standards must be prepared, which are difficult to construct, and are often biased towards the resource that the lexicographer consulted.

In this paper we evaluated the English-Russian dictionary against the following criteria: average number of translation variants in a dictionary entry, the percentage of incorrect translations and the percentage of extremely noisy translations. We used a manually annotated sample of translation equivalents randomly[1] drawn from the dictionary. The annotation task was to determine how well the given translation

[1] Random was used proportionally to the square root of joint frequency, in order to balance rare and frequent phrase pairs in the sample.

equivalent fits for a human-oriented translation dictionary. The annotators classified each translation according to the gradation represented in Table 2.

The average number of translation variants in a dictionary entry is 5.3 per query. A separate experiment has shown that a dictionary entry was found for 97 of 100 random queries to the dictionary.

The evaluation results show that 44.6% of the translation equivalents are unquestionably good, and 41.9% represent the words and phrases that were assessed as being redundant but not incorrect. For example, the dictionary includes many translations of trivial phrases as separate entries (*наша квартира—our apartment, our flat*). The total share of incorrect translations is about 13.5%.

Table 2. The percentage of translation equivalents in the English—Russian dictionary w.r.t. different quality types

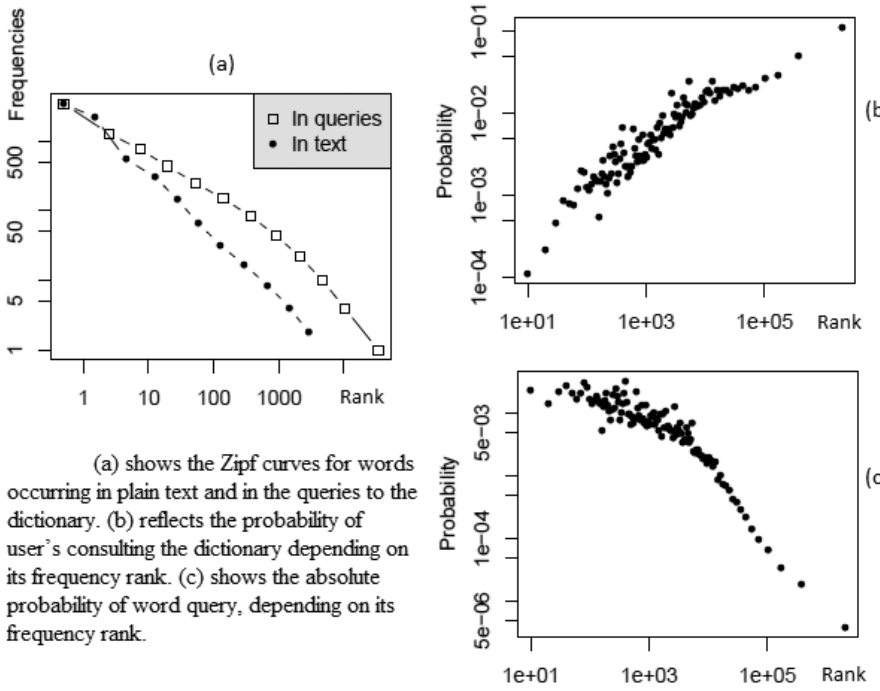
Type	Explanation	% in the dictionary
1	totally wrong or noisy (e.g. misspelled)	3.58
2	incorrect or incomplete translation	9.88
3	not a mistake, but unnecessary translation	41.90
4	good, but not vital	25.21
5	vital translation (must be present in human-built dictionary)	19.43

The evaluation does not take into account the frequency of queries and the order of translations in the dictionary entry. The first translations of frequent words are usually correct.

4.1. Analysis of Dictionary Use

The statistics of dictionary queries is relevant for the analysis of the dictionary quality and for the development of proper evaluation metrics. The properties of dictionary entries for frequent words may differ from those for rare words. Furthermore, the frequencies of words in dictionary queries may differ from their frequencies in text.

In this regard, we conducted an experiment, the purpose of which was to compare the frequency distribution of words in the queries to the dictionary with a uniform distribution, as well as to the frequency distribution of words in texts. We selected one-word queries from the dictionary log for a period of 18 days. Misspelled words were not considered. The frequencies of these words in queries were compared to the frequencies of the same words, collected over a large volume of texts on the Internet. The results are shown in Fig. 4. The points on the X axis correspond to words, ranked in descending order by frequency of occurrence in plain text. We can conclude that the distribution of words in dictionary queries is neither uniform, nor identical to the distribution of words in text. The Zipf curve for the queries decreases more slowly in the area of rare words. The reason is that rare words are likely to be unfamiliar, and the users need to consult the dictionary more often.



(a) shows the Zipf curves for words occurring in plain text and in the queries to the dictionary. (b) reflects the probability of user's consulting the dictionary depending on its frequency rank. (c) shows the absolute probability of word query, depending on its frequency rank.

Fig. 4. Comparison of the distributions of words in queries and text

Conclusion

We have described the procedure for the automatic construction of a large-scale translation dictionary, and the methods for augmenting dictionary entries with useful information, such as context examples, reverse meanings, grouping of synonyms. We also discussed the problem of detection of noisy translations, which is important for the human-oriented dictionary.

The results of the evaluation of the English-Russian dictionary demonstrated the perspectiveness of the overall approach, w.r.t. the coverage of the dictionary, and the depth of its entries. While the noisy translations still occur, their percentage is moderate. We provided the analysis of dictionary use, and discussed the difference between the distributions of words in dictionary queries and plain text.

The lemmatization and filtering ungrammatical phrases require additional morphological and syntactic tools. While the lack of such tools imposes certain restrictions on the dictionary content, the proposed approach is still applicable to many language pairs.

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. *Sue Atkins.* (1994). A corpus-based dictionary. In: Oxford-Hachette French Dictionary (Introductory section). Oxford: Oxford University Press. xix—xxxii
3. *Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer.* (1993). The Mathematics of Statistical Machine Translation: Parameter estimation. Computational Linguistics, 19(2):263–312, June.
4. *Hartmann, R. R. K.* (1994). The use of parallel text corpora in the generation of translation equivalents for bilingual lexicography. In W. Martin, et al. (Eds.), Euralex 1994 Proceedings (pp. 291–297). Amsterdam:Vrije Universiteit.
5. *Philipp Koehn, Franz Josef Och, and Daniel Marcu.* (2003). Statistical phrase-based translation. In Proceedings of HLT-NAACL 2003, pages 127–133.
6. *Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondˇrej Bojar, Alexandra Constantin, and Evan Herbst.* (2007). Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pages 177–180, Prague, Czech Republic
7. *I. Dan Melamed.* (1996). Automatic construction of clean broad-coverage translation lexicons. In Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas, pages 125–134, Montreal , Canada
8. *Franz Josef Och and Hermann Ney.* (2000). Improved Statistical Alignment Models. Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 440–447, Hongkong, China.
9. *Franz Josef Och and Hermann Ney.* (2004). The Alignment Template Approach to Statistical Machine Translation. Computational Linguistics, vol. 30 (2004), pp. 417–449.
10. *Resnik, Philip and Noah A. Smith.* (2003). The web as a parallel corpus. Computational Linguistics, 29, pp. 349–380
11. *Serge Sharoff.* (2004). Harnessing the lawless: using comparable corpora to find translation equivalents. Journal of Applied Linguistics 1(3), 333–350.
12. *Jason Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch and Adam Lopez.* (2013). Dirt Cheap Web-Scale Parallel Text from the Common Crawl. To appear in Proceedings of ACL 2013.
13. *Tomaszczyk, J.* (1986). The Bilingual Dictionary under Review. Snell-Hornby, M. (Ed.). ZuriLEX'86 Proceedings. University of Zurich, Switzerland: 289–297.
14. *Dan Tufis¸ and Ana-Maria Barbu.* (2001). Computational Bilingual Lexicography: Automatic Extraction of Translation Dictionaries. In International Journal on Science and Technology of Information, Romanian Academy, ISSN 1453–8245, 4/3–4, pp. 325–352
15. *Velupillai, Sumithra, Martin Hassel, and Hercules Dalianis.* (2008). "Automatic Dictionary Construction and Identification of Parallel Text Pairs." Proceedings of the International Symposium on Using Corpora in Contrastive and Translation Studies (UCCTS).

ПРОЦЕССЫ ИДИОМАТИЗАЦИИ И ГРАММАТИКАЛИЗАЦИИ В НЕСТАНДАРТНЫХ КОНСТРУКЦИЯХ[1]

Апресян В. Ю. (vapresyan@hse.ru)

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

В статье проводится корпусное исследование конструкции вида *к*-слово+*ни X, P* (*Как ни трудно, надо стараться, Куда ни сунешься — везде отказ*) как представителя особого класса синтаксических объектов — нестандартных конструкций, которое позволяет сформулировать следующие их свойства. Во-первых, большой процент всех встретившихся в корпусе реализаций конструкции составляют сочетания с одной или несколькими лексемами-«фаворитами» (*как ни странно, как ни старался, что ни говори, куда ни глянь* и пр.). Такая же тенденция обнаруживается и у других нестандартных конструкций (*X X-ом, при всем X-е, X-овый X-овый*). Во-вторых, нестандартные конструкции не композициональны — их интерпретации сильно зависят от частеречных и семантических свойств лексем-«фаворитов». Наконец, реализации с некоторыми лексемами-«фаворитами» идиоматизируются и грамматикализуются, «откалываясь» от порождающей конструкции (*куда ни плюнь, как волка ни корми* и т. п.). На выбор лексем-«фаворитов» влияет взаимная аттракция (семантическое согласование) семантики конструкции и семантики заполняющей лексемы, а также особенности языковой картины мира, в которой отражены общие законы мироустройства, представленные в языке.

Ключевые слова: нестандартная конструкция, синтаксическая фразема, валентность, сочетаемость, частотность, лексемы-«фавориты», идиоматизация, грамматикализация

[1] Исследование поддерживалось грантом Программы фундаментальных исследований отделения историко-филологических наук РАН «Язык и литература в контексте культурной динамики», проект «Создание электронной базы данных по Новому объяснительному словарю синонимов русского языка» (2012–2014) и грантом НШ-3899.2014.6 для поддержки научных исследований, проводимых ведущими научными школами РФ «Разработка материалов для Активного словаря русского языка» (2014–2015). В данной научной работе использованы результаты проекта «Корпусные технологии в лингвистических и междисциплинарных исследованиях», выполненного в рамках Программы фундаментальных исследований НИУ ВШЭ в 2014 году.

IDIOMATIZATION AND GRAMMATICALICATION IN NON-STANDARD CONSTRUCTIONS

Apresjan V. Yu. (vapresyan@hse.ru)

National Research University Higher School of Economics, Moscow, Russia

The paper is a corpus research of the Russian construction *wh-word + negative particle X,P* (as in *Kak ni trudno, nado starat'sja* 'However difficult, one has to try'; *Čto on ni prosil, vse emu davali* 'Whatever he asked for, he was given') as a typical representative of a certain class of syntactic objects, namely, non-standard constructions, which reveals the following properties: 1) only one or several lexemes ("favorites") account for up to a half of all encountered realizations; 2) non-standard constructions are non-compositional; 3) realizations with certain "favorites" result in idiomatization and grammaticalization of particular expressions which become separated from the "mother" construction. The choice of "favorites" is triggered by the process of mutual semantic attraction: the interaction of the construction semantics and the semantics of filler lexemes. This choice is also influenced by the linguistic worldview typical of a particular language.

Key words: non-standard construction, syntactic phraseme, valency, combinatory properties, frequency, "favorite" lexemes, idiomatization, grammaticalization

Введение

В данной статье на примере корпусного анализа конструкции вида *к-слово*[2] *+ ни + X, P* (*Как ни старался, ничего не вышло*) рассматриваются общие явления и процессы, связанные с заполнением валентностей нестандартных конструкций, а также их идиоматизацией и грамматикализацией. Несмотря на огромное количество литературы, посвященной конструкциям, в первую очередь, в рамках грамматики конструкций и теории «Смысл-Текст», практически не существует работ, где проводится корпусный анализ частотностей лексемного заполнения переменных в конструкциях (за исключением работ в рамках проекта Rusgram). Представляется, что такой подход дает возможность обнаружить

[2] Под *к*-словами понимаются, как и в работе [Иомдин 2010] местоимения *кто, что, какой, как* и т. д.

явления в области взаимодействия синтаксиса и семантики, а также в области идиоматики, которые не заметны при рассмотрении конструкции на материале нескольких реализаций.

В классификации типов синтаксических конструкций мы опираемся на работу [Иомдин 2010], выделяющую три типа конструкций — стандартные, или конструкции «большого синтаксиса», и конструкции малого синтаксиса, которые делятся на нестандартные конструкции и синтаксические фраземы[3]. Наше понимание различий между этими тремя типами конструкций во многом, но не полностью совпадает с пониманием, представленным в работе Л. Л. Иомдина. Ниже представлены основания для разделения этих трех типов синтаксических объектов.

В существующей лингвистической литературе для классификации фразеологических синтаксических единиц обычно используются следующие параметры: степень композициональности [Fillmore et al. 1988], [Nunberg, Sag & Wasow 1994], [McGinnis 2002], [Mateu and Espinal 2007]; наличие переменных [Fillmore et al. 1988], [Jackendoff 1997]; частотность и степень лексикализации [Иомдин 2010].

Наличие переменных позволяет отличить синтаксические фразеологические единицы от лексических (идиом): и стандартные конструкции, и нестандартные конструкции, и синтаксические фраземы имеют переменные, т. е. варьирующиеся компоненты и, соответственно, не-единственную реализацию.

Стандартные синтаксические конструкции наиболее композициональны и частотны и наименее лексикализованы, однако полная композициональность и отсутствие лексикализации не постулируется даже для них. Ср., например, семантически мотивированные сочетемостные ограничения на заполнение сочинительной конструкции (*X и Y*), где сочиняемые члены X и Y должны иметь некоторое общее основание для сопоставления; так, вне контекста неуместны фразы типа **снег и макароны* (при нормальности фразы, где такое основание создается *Я ненавижу две вещи — снег и макароны*)[4]. Что касается лексикализации, то хотя количество возможных вариантов заполнения сочинительной конструкции никак не ограничено, существуют некоторые устойчивые пары X-ов и Y-ов; ср. *снег и град, студенты и школьники, овощи и фрукты, молодая*

---

[3] Эта типология пересекается, но не совпадает с типологией синтаксических конструкций, предложенной в теории «Смысл-Текст»; ср. [Иорданская, Мельчук 2007: 300–301], где не проводится специального различия между нестандартными конструкциями и синтаксическими фраземами. Под понятие «синтаксической» фраземы, как она выделяется в работе Л. Н. Иорданской и И. А. Мельчука, подпадает семантически некомпозициональная часть «нестандартных конструкций» и «синтаксических фразем» в смысле Л. Л. Иомдина.

[4] Вообще семантические ограничения на сочетания конъюнктов в сочинительной конструкции — это отдельная огромная тема; не пытаясь осветить ее в пределах данной работы, отсылаем читателя к разделам 3.3.1 и Библиография в статье О. Е. Пекелис «Сочинение» в онлайн-проекте rusgram.ru. Здесь упомянем в этой связи лишь работы [Лауфер 1987] и [Санников 1989], где обсуждаются ограничения на заполнение элементов сочинительной конструкции для русского языка.

и красивая, чуткий и внимательный и т. п., где появление в контексте X-а с некоторой степенью вероятности предсказывает появление Y-а[5].

Стандартным конструкциям противопоставлены синтаксические фраземы — сильно лексикализованные единицы, имеющие очень ограниченное число реализаций (обычно в пределах десятка или двух) в силу сильно специализированной семантики и жестких сочетаемостных ограничений на заполнение переменных, которые заполняются лексемами из одного узкого семантического класса. Таковы, например, фраземы *с X-а на X* (*со дня на день, с минуты на минуту, с часа на час*); *X в X (минута в минуту, секунда в секунду, час в час, день в день*); *из X-а в X* (*из конца в конец, из угла в угол, из комнаты в комнату*). Синтаксические фраземы могут иметь разную степень частотности и композициональности, однако заполнение переменных у них высоко лексикализовано. Отличие нашего понимания синтаксических фразем от того, которое представлено в работе [Иомдин 2010] состоит в критерии лексикализации: мы считаем синтаксическими фраземами те конструкции, в которых высоко лексикализовано заполнение переменных, независимо от того, сколько фиксированных лексических элементов есть в самой конструкции (их может и не быть вовсе); в работе [Иомдин 2010] под синтаксическими фраземами понимаются такие единицы, у которых лексически зафиксировано не менее двух элементов, при этом есть переменная (или переменные). Приведем пример: в нашем понимании синтаксической фраземой является конструкция *быть X-у Y-овым Z-ом* (*Она была ей хорошей матерью*), хотя в ней лексически фиксирован только один элемент (связочный глагол *быть*). Однако разумно считать ее фраземой, поскольку заполнение переменных X, Y и Z очень ограничено. Так, прилагательное *Y-овый* может выражать только хорошую или плохую оценку (*прекрасный, отличный, настоящий, верный, плохой, никудышный* и прочие (ANTI) MAGN-ы и (ANTI)BON-ы.

Существительные X и Y — это реляционные имена, симметричные термины близкого родства или постоянных личных отношений *(муж-жена, брат-сестра, ребенок-родитель, спутник-спутница, друзья, подруги)*; ср. *Он был ей хорошим мужем <сыном, братом>, Она была ему хорошей женой <дочерью, спутницей>, Она была мне верной подругой, Он был мне прекрасным другом*, но не **Она была им профессиональной учительницей*; **Она была ему компетентной сотрудницей*; **Он был нам прекрасным заведующим*; **Он был хорошим любовником жене своего друга.* С другой стороны, такие единицы как *мало + к-слово X (мало кто, мало что, мало какой* и т. п.) мы считаем нестандартными конструкциями, хотя в них лексически фиксировано два элемента, поскольку они позволяют практически неограниченное заполнение переменной X.

Итак, нестандартные конструкции занимают промежуточное положение: с одной стороны, их значение и, соответственно, сочетаемость существенно уже, чем у стандартных конструкций, с другой стороны, они имеют намного более

---

[5] Об устойчивых сочинительных сочетаниях такого рода см. раздел 4.1 в статье О. Е. Пекелис «Сочинение» в онлайн-проекте rusgram.ru, где в этой связи упоминаются, в том числе, работы Лауфер [1987], [Санников 1989] и [Урысон 2011].

широкие значения и значительно более гибкую сочетаемость, чем синтаксические фраземы, т. е. лексическое заполнение их переменных потенциально бесконечно и ограничено лишь очень широкими семантическими классами.

Корпусное исследование частотностей и семантики различных реализаций (вариантов заполнения валентностей) нестандартной конструкции *к*-слово + *ни X, P* обнаруживает интересные закономерности, которые позволяют сформулировать гипотезы о некоторых общих лингвистических тенденциях: 1) о механизмах образования некоторых типов идиом; 2) о механизмах грамматикализации; 3) о правилах семантического согласования между синтаксическими и лексическими единицами; 4) об общих законах мироустройства, представленных в языке.

1. Реализации и интерпретации конструкции *к*-слово + *ни X, Y* с разными *к*-словами

В конструкции *к*-слово + *ни X, P* есть три вариативных элемента — *к*-слово, X и P: *Как [к-слово] мы ни уговаривали [X] его, он отказался [P]*; *Сколько [к-слово] мы с ним ни возились [X], толку не вышло [P]*.

Данная конструкция реализуется со всеми *к*-словами (*как, что, сколько, куда, кто, какой, где, когда, откуда*), кроме *зачем*. Три встретившиеся в корпусе примера с *зачем* содержат частицу *бы* и, соответственно, представляют собой реализации другой, хотя и близкой, конструкции *к-слово + бы + ни X, P* и здесь не рассматриваются: *Зачем бы он ни явился сюда, он здесь, перед нею* (Н. Шпанов).

Реализации этой конструкции представлены пятью типами единиц, между которыми не всегда существуют строгие границы:

а) Грамматикализованные единицы (серия неопределенных местоимений на *-нибудь*);
б) Идиомы (*Куда ни кинь, везде клин*);
в) Устойчивые коллокации (*Как он ни старался, ничего не вышло*);
г) Свободные сочетания (*Как жена ни расспрашивала, он ничего ей не сказал*);
д) Coercion[6], или «давление синтаксиса на семантику»[7] (*Как она ни копала налоговую отчетность, все было в полном порядке*).

---

[6] Мы используем этот термин так, как он используется в работах по грамматике конструкций, например [Goldberg 1995]: когда семантика конструкции «вынуждает» предикат интерпретироваться определенным образом, хотя изначально это предикат не подходит для этой конструкции по своим семантическим свойствам. Глагол *копать* не градуируем; нельзя **очень копать*,**немного копать* (хотя возможно *глубоко копать*). Однако семантика конструкции «навязывает» ему интерпретацию высокой степени: *как ни копала* значит ‘долго, тщательно, прилагая большие усилия’.

[7] Вообще представление о том, что предикат может «форсированно» менять значение, будучи помещен в определенные синтаксические условия, было введено существенно раньше в работе [Апресян 1967: 29–31] на примере глаголов физического

Хотя в данной конструкции возможны практически все вопросительные местоимения, частотности реализаций с разными местоимениями резко отличаются — как будет показано, в силу семантических особенностей конструкции. Ниже приводятся частотности реализаций фраземы с разными вопросительными местоимениями в Основном корпусе НКРЯ (при запросе вопросительное местоимение, *ни* на расстоянии 1, предикат на расстоянии 1).

Таблица 1

Как ни	9510
Что ни	2563
Куда ни	1112
Сколько ни	857
Кто ни	564
Какой ни	357
Откуда ни	309
Где ни	250
Когда ни	228

Конструкция имеет две основных интерпретации, в зависимости от *к*-слова и типа предиката, заполняющего валентность X:

(1) **Интерпретация «любого варианта»**
‘при любом варианте X имеет место P’
Кто ни приходил, она всем помогала
Как ни расставляли мебель, получалось некрасиво

(2) **Уступительная интерпретация**
‘X имеет место в высокой степени; говорящий считает, что обычно, если X, имеет место ситуация не-P; имеет место P’
Каким он ни был подлецом, она его любила
Сколько я ему ни объяснял, он все равно ничего не понял

Семантически уступительная интерпретация представляет собой сужение интерпретации «любого варианта» в применении к градуируемым предикатам. С градуируемыми предикатами сочетаются только местоимения, вводящие указание на шкалу: *как* (в значении ‘в высокой степени’ или ‘интенсивно’), *какой* (в значении ‘в высокой степени’), *сколько* в значении ‘долго’ или ‘интенсивно’; таким образом, чисто уступительная интерпретация возможна только для реализаций *как ни, какой ни, сколько ни*. Для *как ни* и *сколько ни* она является подавляюще частотной.

воздействия типа *валять, драть, дуть, жарить, катать, резать, садить, сыпать, хватить, чесать*, которые могут приобретать значения 1) типа «бить», «ударять»; 2) значение типа «идти», «бежать»; 3) значение типа «говорить», «писать»; 4) значение типа «играть», «плясать».

Из таблицы 1 видно, что реализации *как ни* и *сколько ни* более чем в три раза превосходят по частотности все остальные, что значит, что уступительная интерпретация является предпочтительной для данной конструкции. Для того, чтобы ответить на вопрос, почему это так, необходимо объяснить, каким образом происходит переход от семантики «любого варианта» к семантике уступительности.

Главная семантическая идея интерпретации «любого варианта» — отсутствие разницы между всеми возможными вариантами X с точки зрения наступления ситуации P; главная идея уступительного значения — что имеющий место неблагоприятный для P вариант X не оказывает ожидаемого воздействия и P имеет место.

Семантически эти идеи не идентичны: ср. ‘произвольно выбранный из всех возможных X’ vs. ‘такой X, который препятствует P’. Однако прагматически сочетания с квантором *любой* типа *в любом случае P, при любых обстоятельствах P, в любое время P* используются, когда необходимо подчеркнуть, что P будет иметь место даже при самых экстремальных вариантах — в самых неблагоприятных обстоятельствах, в самое неудобное время и т. д. Ср. также идиому *любой ценой*, которая имеет единственную интерпретацию ‘самой дорогой ценой, самыми трудными способами’, но не ‘самой дешевой ценой, самыми легкими способами’. Трансформацию ‘любой’ => ‘крайний, неблагоприятный, экстремальный’ легко объяснить максимой информативности: отмечать, что P имеет место при благоприятствующем варианте X, неинформативно.

Таким образом, семантическое развитие этой конструкции в сторону уступительности естественно, что и объясняет сильное преобладание реализаций с *как*: уступительная интерпретация задает идею высокой степени и, соответственно, требует градуируемых предикатов, которые вводятся шкалярным местоимением *как*. Более того, даже реализации с нешкалярными местоимениями окрашиваются уступительной количественной семантикой (ср. *Что ни делай, все без толку* = ‘Даже если делать много разного, толку не будет’). В реальном употреблении половина нешкалярных реализаций окрашена уступительной и степенной семантикой; в целом уступительные интерпретации составляют примерно 85 процентов от всех возможных реализаций этой конструкции.

Остается вопрос — почему при переходе от идеи ‘любого варианта’ к идее уступительности из двух возможных полюсов выбирается именно полюс высокой степени, т. е. почему конструкция *к-слово + ни X, P* имеет интерпретацию ‘X имеет место **в очень высокой** степени’, но не ‘X имеет место **в очень низкой** степени’; ср. возможность (5а), но не (5б):

(3) а. *Как он ни просил (= много), она ему отказала*
б. **Как он ни просил (=мало), она согласилась*

В принципе местоимение *как* способно развивать и значение высокой степени *(Как$^{\uparrow}$ они убрали в квартире — все сияет* = ‘очень много и хорошо’*)*, и значение малой степени (*Да как$^{\downarrow}$ они убрали — все так и валяется* = ‘плохо, мало’). Уступительное значение также совместимо с обоими полюсами, как показывает возможное заполнение валентностей союза *хотя:*

(4) а. *Хотя он очень ее просил, она ему отказала*
б. *Хотя он ее не просил, она согласилась*

Однако нестандартная конструкция *к-слово +ни* «выбирает» из двух возможных только полюс высокой степени. Представляется, что это является языковым отражением представлений о законах мироустройства, а именно:

(5) Осуществлению события **мешают большие** препятствия, но **не помогают маленькие** препятствия, так как события прототипически каузируются **активными усилиями**, а не **слабым противодействием**

Уступительные средства фиксируют нарушения естественных ожиданий: соответственно, существование ситуации несмотря на большие препятствия фиксируется уступительными средствами, а не-существование ситуации несмотря на маленькие препятствия — нет. Почему же для союза *хотя* возможны оба типа употреблений? По-видимому, более идиоматичные нестандартные конструкции отражают специфические закономерности мироустройства в более обязательном порядке, чем более композициональные стандартные конструкции типа уступительной конструкции с союзом *хотя*. При этом и для конструкции с *хотя* более частотными будут реализации типа (6а); однако реализации типа (6б) не являются запрещенными.

2. Реализации и интерпретации конструкции *к*-слово + *ни X, Y* с разными предикатами X

В пределах данной работы невозможно подробно проанализировать все реализации конструкции, поэтому будет подробно рассмотрена наиболее частотная реализация *как ни* и кратко сформулированы основные результаты частотного исследования прочих реализаций. Больше половины всех реализаций *как ни* составляет реализация с прилагательными (5229 в Основном корпусе), что неудивительно, т. к. в основном сочетания с *как ни* интерпретируются уступительно, уступительная интерпретация возможна с градуируемыми предикатами, а признаки — это классические градуируемые сущности; ср. *Как ни соблазнительно было предложение, мы его отвергли*, *Как ни красива она была, она его совершенно не привлекала*.

2.1. Реализации как ни + Adj

Анализ частотностей разных прилагательных, заполняющих валентность X, дает следующую картину: примерно половину всех реализаций (2227 в Основном корпусе) составляет прилагательное *странный* в составе оборота *как ни странно*; другие достаточно частотные реализации — это синонимы

странно типа *удивительный* и *парадоксальный* в составе оборотов *как ни удивительно, как ни парадоксально* (314), *тяжелый* (113) и *трудный* (57), а также прилагательные *мал* (110), *велик* (84), *хороший* (55), *плохой* (47), *прискорбный* (45), *грустный* (43). На прочие реализации приходится 2158 вхождений, причем ни одна из них по частотности не превышает 0.5 процента:

Таблица 2

как ни странно	2227
как ни удивительно, *как ни парадоксально*	314
как ни тяжело *как ни трудно*	170
мал	110
велик	84
хороший	55
плохой	47
как ни прискорбно	45
как ни грустно	43
ПРОЧЕЕ	2158

Обращает на себя внимание сразу несколько фактов. Во-первых, у конструкции существует лексема-«фаворит», которая составляет практически половину всех ее реализаций. Этот лингвистический феномен напоминает принцип Парето (эмпирическое правило, изначально сформулированное для экономических и социологических явлений), согласно которому «20% усилий дают 80% результата, а остальные 80% усилий — лишь 20% результата» (реальные числа могут быть иными, смысл правила в том, что большая часть какой-либо деятельности обычно реализуется меньшей частью активных участников). В применении к заполнению нестандартных конструкций — большой процент реализаций образуется малым числом лексем-«фаворитов». Во-вторых, «топовые» реализации относятся к семантическим классам, которые являются классическими модификаторами уступительного значения — вероятность (*странно, удивительно, парадоксально*), степень (*мал, велик*), (не)желательность (*хороший, плохой, прискорбный, грустный*)[8], т. е. имеет место семантическое согласование. В-третьих, большинство «топовых» реализаций — это не свободные реализации, а устойчивые вводные обороты, идиоматизированные и грамматикализованные.

Проиллюстрируем сказанное на примере вводного оборота *как ни странно*; ср.

***Как ни странно**, наиболее правдоподобным выглядит объяснение президентской пресс-службы* («Еженедельный журнал», 2003.03.17. Оборот *как ни странно* представляет собой идиоматизированную и грамматикализованную единицу. Во-первых, *как ни странно* не представляет из себя полноценной зависимой

[8] О семантических модификациях уступительного значения см. [В. Апресян 2006].

клаузы, а является типичным вводным оборотом, который может вставляться в практически любое место в предложении — в начало, середину или конец главной клаузы, — что затруднено для не-идиоматизированных реализаций типа *Как ни соблазнительно было предложение, мы его отвергли*. Во-вторых, в этом обороте не упоминается объект, чья странность оценивается, т.е. отсутствует полноценное выражение валентности X. По сути, оборот представляет из себя катафорическую или анафорическую единицу с эллиптированным дейктическим местоимением *это*, кореферентным ситуации P, описываемой в главной клаузе: *как (это) ни странно, P.* Таким образом, семантика конструкции *к-слово +ни + X, P* редуцируется в данной реализации до модальной рамки:

(6) *Как ни странно (P), P* = 'имеет место P; говорящий считает, что это очень странно'

Выражение *как ни странно, P* отражает следующий закон мироустройства:

(7) 'Странность ситуации является препятствием к ее существованию'

Другие «топовые» идиоматизированные реализации отражают другие законы мироустройства. Ср. оборот *как ни тяжело признавать <признаваться, сознавать, смириться>, P*: *Ошибки нужно исправлять,* ***как ни тяжело*** *это сознавать* («Воздушно-космическая оборона», 2003.04.15); *«Существование» — это только модус человеческого мышления,* ***как ни тяжело*** *с этим смириться* (И. Савельева, А. Полетаев); *Но у меня —* ***как ни тяжело*** *было тогда в том признаваться — статья эта породила, наоборот, протест и глубокую тревогу* (А. Борин), со следующим значением, также состоящим из модальной рамки:

(8) *Как ни тяжело Y-у сознавать <признавать, смириться с> P, P* = 'имеет место P; говорящий считает, что У-у неприятно P, и что поэтому Y-у тяжело сознавать, что P <признавать P, смириться с P>'

Этот оборот отражает крайнюю степень солипсизма в качестве закона мироустройства:

(9) 'Препятствием к существованию ситуации является трудность ее осознания субъектом'

Обороты *как ни грустно, как ни прискорбно* фиксируют нежелательность ситуации в качестве препятствия для ее существования.

В реализации с прилагательным *мал* препятствием к существованию ситуации P выступает очень малая степень этого существования (*Как ни мала вероятность этого события, она существует*; *Как ни мала эта величина, нулю она не равняется*). Опять-таки, как и в случае *как ни странно, как ни грустно, как ни тяжело* в качестве препятствия к существованию P выступает не какая-то другая ситуация, а определенная оценка P, т.е. фразема в этой реализации также

представляет из себя модальную рамку. Закон мироустройства, отраженный в данной реализации фраземы, может быть сформулирован как

(10) ‘Препятствием к существованию ситуации является ее малая степень’.

Реализации с антонимом *великий* (*Как ни велик X, P*) устроены совершенно асимметрично реализациям с *малый*, а именно, они представляют из себя классические примеры уступительного значения, где *велик* модицифирует общеуступительное значение по степени:

(11) *Как ни X, P* = ‘имеет место X в очень высокой степени; имеет место P; говорящий считает, что обычно, если имеет место ситуация типа X, то имеет место ситуация типа не-P’.
Как ни велик был соблазн вставить кассету в диктофон [X] — он лежал на рабочем столе Маревой, но Страхов его преодолел [P] (Газета (2000))

Существенная часть реализаций *как ни плохо X, P* и антонимичным им реализаций *как ни хорошо X, P* выражает оценку — а именно, хотя ситуация или объект P квалифицируются как очень плохие (очень хорошие), говорящий утверждает, что у P есть какие-то достоинства (недостатки) или что есть другие ситуации или объекты, которые еще хуже (лучше). Таким образом, эти реализации (близкие по частотности) отражают два симметричных постулата мироустройства — один оптимистичный (14), другой пессимистичный (15):

(12) ‘Всегда можно найти что-то худшее; даже у самого плохого могут быть хорошие стороны’
Как ни плохо *дома, все же вы у себя, среди своих* (В. Теляковский, А. Южин)
Как ни плоха *шинель для спанья, голые нары — хуже* (В. Войнович)

(13) ‘Всегда можно найти что-то лучшее; даже у самого хорошего могут быть плохие стороны’
Как ни *хороша наша новогорская база, осточертели эти четыре стены* (Е. Рубин).
Как ни *хороша была Строцци, но Трутовская в этой группе была самой сильной балериной* (Л. Лопато).

Заканчивая рассмотрение реализации *как ни + Adj*, ответим, почему именно *странный* составляет львиную долю реализаций конструкции *как ни X, P*. Представляется, что это происходит в силу взаимной аттракции уступительной семантики конструкции, предназначенной для того, чтобы фиксировать странные, необычные положения вещей, и семантики прилагательного: смысл прилагательного дублирует смысл конструкции. Что касается того, почему из синонимического ряда *странный, необычный, удивительный, поразительный, парадоксальный* и пр. доминирует именно *странный*, то ответ, по-видимому, таков: из тех прилагательных, что способны к предикативному

употреблению в качестве вводных оборотов (*странно, удивительно, поразительно, парадоксально*) *странный* является доминантой и по семантическим свойствам, и по частотности (20579 вхождений *странно* в Основном корпусе vs. 10387 вхождений *удивительно* vs. 1882 вхождения *поразительно* vs. 673 вхождения *парадоксально*). Интересно, что при попадании в конструкцию *как ни* численное превосходство прилагательного *странный* над его синонимами вырастает экспоненциально.

2.2. Реализации как ни + V

Реализации *как ни* с глаголом представляют собой другую семантическую модификацию рассматриваемой конструкции. Большая часть реализаций имеет уступительную интерпретацию, с *как* в значении показателя высокой интенсивности или степени. Имеется и некоторое количество интерпретаций ‘любого варианта’, где *как* имеет значение способа: *Буквально каждое слово,* ***как ни подай*** *его, будет подхвачено партнером* (А. Дмитриев); *Что ни скажет, — все хорошо;* ***как ни*** *ступит, — все ловко* (П. Ю. Львов). При этом даже интерпретации ‘любого варианта’ имеют уступительную окраску. Для них предлагается следующая модификация значения:

(14) *Как ни X, P* = ‘ситуация P будет иметь место при любом варианте выполнения действия X; говорящий считает, что обычно при некоторых вариантах выполнения действия X имеет место ситуация типа не-P’.

Интерпретация конструкции *как ни X, P* в значении ‘любого варианта’ vs. в уступительном значении коррелирует с формой (видом, временем и наклонением) глагола, заполняющего валентность X. В целом интерпретация «любого варианта» тяготеет к формам СОВ, а уступительная — к формам НЕСОВ. Это имеет семантическое объяснение: уступительность градуирует, задавая высокую степень интенсивности действия, в то время как интерпретация «любого варианта» задает перебор возможных вариантов осуществления действия. Интенсивные (и вообще градуируемые по интенсивности) действия прототипически выражаются формами НЕСОВ (*Он очень старается, Он ее сильно уговаривал*), в то время как с формами СОВ интерпретация интенсивности затруднена из-за того, что СОВ свойственно представлять действие не как континуум, а пунктивно. Кроме того, уступительные интерпретации чаще фактивны, а интерпретации ‘любого варианта’ чаще потенциальны, что влияет и на выбор времени и наклонения. В соответствии с семантикой грамматических форм и семантикой конструкции формы и интерпретации «выбирают» друг друга — формы НЕСОВ (обычно в прошедшем времени и изъявительном наклонении) дают в основном уступительную интерпретацию, формы СОВ (обычно в будущем времени или императиве) — интерпретацию «любого варианта»; ср. *Но розового куста, выращенного Хансом,* ***как ни смотрели*** *в промзоне, мы не увидели* (А. Приставкин) [уступительная интерпретация] vs. *На фотографии была*

другая девушка. ***Как ни посмотри*** (Э. Лимонов) [интерпретация «любого варианта»]. Поскольку в реализации *как ни* доминирует уступительная интерпретации, имеется четкая тенденция к доминированию форм НЕСОВ: *как ни* + НЕСОВ — 3743 вхождений в Основном корпусе, *как ни* + СОВ — 644.

Рассмотрим уступительные интерпретации конструкции. В силу категориальной семантики глагола как прототипического действия, основной массив реализаций *как ни* + V моделирует ситуации, где, несмотря на большие усилия агенса, направленные на осуществление Р, Р не осуществляется; ср. *Как он ни напрягал зрение, он ничего не увидел*:

(15) *Как ни X, P* = 'Агенс очень интенсивно или очень долго прилагает усилия X, чтобы осуществилась ситуация Р; ситуация Р не имеет места; говорящий считает, что обычно если прилагать усилия X, то имеет место ситуация типа Р'.

Среди глаголов, заполняющих валентность X, также определяются четкие «фавориты». Среди 3743 вхождений НЕСОВ в Основном корпусе частотности глаголов распределились следующим образом:

Таблица 3

стараться	562
как ни крути	310
биться	94
пытаться	70
уговаривать	54
как ни верти	50
как волка ни корми	25
ПРОЧЕЕ	2578

Как видно из таблицы, 15 процентов от всех употреблений конструкции *как ни* в НЕСОВ составляют устойчивые коллокации с глаголом *стараться* (562 вхождения), поскольку он является прототипическим обозначением усилия и, таким образом, наиболее семантически подходящим кандидатом для данной реализации конструкции; высока также доля его синонимов *биться* (94) и более «слабого» *пытаться* (70), а также иллокутивов типа *уговаривать* (54). Особый подтип этой интерпретации представляют идиомы — *как ни крути* (310) и ее синоним *как ни верти* (50), где уступительное значение трансформируется в модальную рамку, близкую по смыслу оборотам *как ни тяжело это признать, как ни прискорбно* и т. п.; ср. *Это ведь, как ни крути, воровство* (М. Сергеев). Ср. предлагаемую модификацию толкования:

(16) 'Говорящий считает, что ситуация Р нежелательна и что поэтому адресат может рассматривать ситуацию как не-Р; говорящий считает, что при любом способе рассмотрения можно утверждать, что ситуация Р имеет место'.

Еще одна идиома в этой реализации — *как волка ни корми (все равно он в лес смотрит)*, где значение интенсивности трансформируется в значение количества (25 вхождений). Итак, хотя глагольные реализации *как ни* несколько менее лексикализованы, чем реализации с прилагательными, тем не менее треть всех употреблений приходится на устойчивые коллокации или идиомы с несколькими лексемами-«фаворитами», т. е. наблюдается та же тенденция, что и с прилагательными.

3. Лексемы-«фавориты» и идиомы в других реализациях конструкции *к-слово + ни X, P*

Реализации конструкции с другими *к*-словами демонстрируют те же тенденции — большой процент реализаций «задействует» маленькое количество лексем-«фаворитов», реализации с лексемами-«фаворитами» представляют собой устойчивые коллокации или оценочные идиоматизированные и грамматикализованные обороты с редуцированной валентностью X. Выбор предикатов-«фаворитов» в устойчивых коллокациях регулируется семантической аттракцией; для данной конструкции, в силу ее уступительного значения, это — конативы и иллокутивы, т. е. слова со значением попытки и усилия.

В реализации *что ни X, P* «фаворитом» является лексема *говорить* (*сказать*), которая в составе идиоматического вводного оборота *что ни говори* составляет 30 процентов (763) от всех встретившихся в Основном корпусе вхождений *что ни* (2563). Ее значение близко значениям других идиоматичных реализаций этой конструкции, особенно *как ни крути*; ср. *Паршивая штука — старость. Что ни говори, паршивая* (И. Муравьева). Оно также состоит из модальной рамки:

(17) ‘Говорящий считает, что ситуация P нежелательна и что поэтому адресат может привести возражения против того, что P имеет место; говорящий считает, что при любых возражениях можно утверждать, что ситуация P имеет место’.

В реализации *сколько ни X, P* (всего 1291 вхождений) «фаворитом» является конатив *стараться* (101) и другие конативы — *пытаться, биться* и пр. (137), а также иллокутивы — *говорить, уговаривать* и пр. (91), которые образуют устойчивые коллокации с общим для уступительных глагольных реализаций значением:

(18) *Сколько ни X, P* = ‘Агенс очень интенсивно или очень долго прилагает усилия X, чтобы осуществилась ситуация P; ситуация P не имеет места; говорящий считает, что обычно если прилагать усилия X, то имеет место ситуация типа P’.

Вместе с пословицей *сколько волка ни корми (он все равно в лес смотрит)* (22 вхождения), «фавориты» составляют около трети от всех реализаций *сколько ни*.

Наконец, в реализации *куда ни* (1112 вхождений) «фавориты» — идиомы *куда ни кинь (везде клин)* (125), *куда ни плюнь* (37), а также идиоматизированные обороты *куда ни посмотри* с другими вариантами глаголов целенаправленного визуального восприятия (303) и *куда ни пойдешь* с другими вариантами глаголов перемещения (249) — составляют 64 процента от всех реализаций. Все они имеют редуцированную валентность X и общее значение ‘везде’; ср. примеры *Куда ни плюнь, или пивной ларек, или круглосуточный магазинчик, забитый паленой водкой* (М. Хайруллин); *Куда ни посмотри, всё одно железо и камень!* (Ю. Домбровский); *Одна беда: куда ни пойдешь — везде высокие дощатые заборы правительственных дач* (М. Поповский). Добавляется также отрицательный оценочный компонент значения:

(19) ‘Нежелательная ситуация P будет иметь или имеет место во всех местах; говорящий считает, что обычно или хорошо, чтобы в некоторых местах имела место ситуация типа не-P’

Общий закон мироустройства, который отражен в этой интерпретации ‘обычно в разных местах наблюдается разное положение вещей; если одно и то же положение вещей наблюдается везде — это необычно или плохо’.

4. Заключение

Представляется, что приведенные в работе данные по поводу неоднородного заполнения нестандартной конструкции *к-слово + ни X, P*, с идиоматизацией и грамматикализацией наиболее частотных сочетаний, составляющих львиную долю всех реальных реализаций конструкции в узусе, можно обобщить. В работе выдвигается гипотеза о том, что это явление носит достаточно универсальный характер для нестандартных конструкций, позволяя выделить этот тип лингвистических объектов в отдельный класс. Приведем другие примеры того же явления на материале некоторых других нестандартных конструкций. Существует уступительная конструкция с повтором существительного вида *X X-ом, а Y (Y-ом)* [Булыгина, Шмелев 1997], [В. Апресян 2006], [Иомдин 2010], со смыслом ‘Хотя говорящий признает важность X-а, Y тоже важен; Y не зависит или не должен зависеть от X-а’; ср. *Работа работой, а отдыхать тоже нужно*; *Воспитание воспитанием, а генетика генетикой*. «Фаворитом» заполнения этой конструкции служит идиома *Дружба дружбой (а служба службой/а табачок врозь)* (72 вхождения в Основном корпусе), в которой сталкиваются два важных (в русской языковой картине мира) явления, которые в идеале не должны влиять друг на друга, но в реальности, по-видимому, влияют — дружеские и рабочие отношения между людьми. Интересно, что в английском языке нет ее полноценного фразеологического соответствия; вариант перевода, приводимый в словаре Любенской [Lubensky 2004] — *business and friendship don't mix* — отсутствует в корпусе COCA, хотя есть в Google — но в 15 раз реже, чем фразема *дружба дружбой*. Эти данные приводятся в качестве подтверждения тезиса

о том, что приобретение тем или иным заполнением ранга «фаворита» и последующая идиоматизация сочетания с ним коррелирует не только с семантикой конструкции и лексем-заполнителей, но и с языковой картиной мира.

Есть и другие примеры нестандартных конструкций с идиоматизацией заполнений-«фаворитов»; ср. уступительную конструкцию *при всем X-е*, 30 процентов всех реализаций которой составляют три идиомы — *при всем том, при всем желании, при всем уважении*. Итак, обобщая, можно сделать следующие выводы из данного корпусного исследования: (1) некоторые идиомы и грамматические единицы изначально представляют собой частотных «фаворитов» заполнения нестандартных конструкций, «обслуживающих» большой процент реализаций конструкции; (2) на выбор «фаворитов» влияет взаимная аттракция (семантическое согласование) семантики конструкции и заполняющей лексемы, а также особенности языковой картины мира, в которой отражены общие законы мироустройства, представленные в языке.

Литература

1. *Апресян Ю. Д.* Экспериментальное исследование семантики русского глагола. М., 1967.
2. *Апресян В. Ю.* Уступительность в языке // Лингвистическая картина мира и системная лексикография. Под ред. Ю. Д. Апресяна. М., 2006. с. 615–712.
3. *Булыгина Т. В., Шмелев А. Д.* Языковая концептуализация мира (на материале русской грамматики). М., 1997.
4. *Иомдин Л. Л.* Синтаксические фраземы: между лексикой и синтаксисом // Теоретические проблемы русского синтаксиса: взаимодействие грамматики и словаря. Под ред. Апресян Ю. Д. Языки славянских культур. М., 2010. с. 141–190.
5. *Иорданская Л. Н., Мельчук И. А.* Смысл и сочетаемость в словаре. Языки славянских культур. М., 2007.
6. *Лауфер Н. И.* Линеаризация компонентов сочинительной конструкции. В: Кибрик А. Е., Нариньяни А. С. (ред.). Моделирование языковой деятельности в интеллектуальных системах. М.: Наука. 1987. с. 167–176.
7. *Пекелис О. Е.* Сочинение. Материалы для проекта корпусного описания русской грамматики (http://rusgram.ru). На правах рукописи. М., 2013.
8. *Санников В. З.* Русские сочинительные конструкции. Семантика. Прагматика. Синтаксис. М., 1989.
9. *Урысон Е. В.* Опыт описания семантики союзов. Языки славянских культур. М., 2011.

References

1. *Apresyan V. Yu.* Ustupitel'nost' v jazyke [Concession in language]. Lingvisticheskaja kartina mira i sistemnaja leksikografija [Linguistic worldview and systemic lexicography]. Apresyan Yu. D. (ed.). Moscow, 2006. pp. 615–712.
2. *Bulygina T. V., Shmelev A. D.* Jazykovaja kontseptualizatsija mira (na materiale russkoj grammatiki). [Linguistic conceptualization of the world (on the material of the Russian grammar]. Moscow, 1997.
3. *Iomdin L.* Sintaksičeskie frazemy: meždu leksikoj i sintaksisom [Syntactic phrasemes: between lexicon and syntax]. In Teoretičeskie problemy russkogo sintaksisa : vzaimodejstvie grammatiki i slovarja, Apresjan Ju. (ed.). [Theoretical issues in Russian syntax: interaction of grammar and dictionary]. Jazyki slavjanskix kul'tur. Moscow, 2010. pp. 141–190.
4. *Iordanskaja L., Mel'čuk I.* 2007. Smysl i sočetajemost' v slovare. [Meaning and co-occurrence in dictionary]. Jazyki slavjanskix kul'tur. Moscow.
5. *Espinal T., Mateu J.* Argument Structure and Compositionality in Idiomatic Constructions. The Linguistic Review. 24:1. 2007. 33–59.
6. *Fillmore C. J., Kay P. & M. C. O'Connor.* 1988. Regularity and Idiomaticity in Grammatical Constructions: the Case of Let Alone. Language. 502 (64, 3). 501–538.
7. *Goldberg, A.* Constructions: A Construction Grammar approach to argument structure. Chicago: University of Chicago Press, 1995.
8. *Jackendoff, R.* 1997. Twistin' the Night Away. *Language* 67. 320–338.
9. *Laufer N. I.* Linearizacija komponentov sočinitel'noj konstrukcii [Linearization of components of coordinative conjunction]. V: Kibrik A. E., Narin'jani A. S. (eds.). Modelirovanije jazykovoj dejatel'nosti v intellektual'nyx sistemax [Modeling of linguistic activity in intellectual systems]. Nauka. Moscow. 1987. p. 167–176.
10. *Lubensky S.* 2004. Random House Russian-English Dictionary of Idioms. New York.
11. *McGinnis, M.* On the Systematic Aspect of Idioms. *Linguistic Inquiry* 33:4. 2002. 665–672.
12. *Nunberg, G., Sag I., Wasow T.* Idioms. *Language* 70. 1994. 491–538. Pekelis O. E. Sočinenije. [Conjunction]. Materialy dlja proekta korpusnogo opisanija russkoj grammatiki (http://rusgram.ru). [Materials for the project of corpus description of Russian grammar]. Moscow, 2013.
13. *Sannikov V. Z.* Russkie sočinitel'nye konstrukcii. Semantika, Pragmatika. Sintaksis. [Russian coordinative constructions. Semantics. Pragmatics. Syntax]. Moscow. 1989.
14. *Uryson E. V.* Opyt opisanija semantiki sojuzov. [At attempt at describing the semantics of conjunctions]. Jazyki slavjanskix kul'tur. Moscow. 2011.

AUTOMATIC ENRICHMENT OF INFORMAL ONTOLOGY BY ANALYZING A DOMAIN-SPECIFIC TEXT COLLECTION[1]

Astrakhantsev N. A. (astrakhantsev@ispras.ru),
Fedorenko D. G. (fedorenko@ispras.ru),
Turdakov D. Y. (turdakov@ispras.ru)

Institute for System Programming of the Russian Academy of Sciences, Moscow, Russia

The core part of an entity linking system, in particular one oriented to wikification, is ontology, which is often informal and supports semantic relatedness as the only type of relation. Most of these systems suffer from the problem of ontology incompleteness. It is especially important for specific domains, since often the only source of extractable knowledge is plain text. This paper formulates the incompleteness problem as a task of ontology enrichment from domain-specific texts and presents a novel approach that combines state-of-the-art methods for terminology enrichment, our own ML-based method for homonymy detection, and methods adopted from the related field for relations extraction. Experimental evaluation shows that the bottleneck is terminology enrichment step: its average precision is about 35%, which is inapplicable for automatic usage, especially taking into account the strict requirements for ontology correctness; however, recall is high enough to help semi-automatic terminology enrichment. We also show that the best features for terminology enrichment differ from those for classic terminology recognition task.

Key words: ontology enrichment, terminology recognition, terminology enrichment, knowledge base construction, entity linking, wikification

1. Introduction

Transition from words to their meanings is essential for many natural language processing applications [2]. An important and extensively researched example is *wikification*—"the task of identifying concepts and entities in text and disambiguating them into their corresponding Wikipedia page" [4]. Some authors call this task word sense disambiguation (WSD), others prefer *entity linking* and specify concepts as "meaningful entities that have properties, semantic types, and relationships with each other" [19]. From the last definition it is obvious that in order to perform such entity linking, one should have a set of concepts with relations between them, what constitutes ontology, or knowledge base[2]. Worth noting, entity linking is not the only

[1] The reported study was supported by RFBR, research project No. 14-07-00692

[2] Some authors consider knowledge base to be a set of concepts, while ontology is "a schema for knowledge base" [26]. However, we have found term *ontology* to be commonly used in both meanings, especially in terms 'ontology learning' or 'ontology enrichment'

application of ontologies, they are widely used in Question Answering [36], Information Retrieval [14, 16], and so on.

All ontology based systems share the problem of incompleteness: even for a narrow domain with available hand-crafted ontology, there are missing concepts due to domain evolution. Moreover, domain knowledge is usually encoded in collections of plain texts only. Entity linkers can approach this problem two ways: (a) during the document processing, detect words that currently have no concept in the ontology (e.g. [19]); (b) extract new concepts with relations from domain-specific texts as a separate activity and enrich the ontology by this data.

The presented paper follows the second way, or ontology enrichment (OE), for specific domains. We preferred it to the first one for the following reasons:

- We can extract more knowledge about concepts, since we have more data about their occurrences;
- Entity linking algorithms keep simpler and faster, thus OE can be more resource consuming.

Our task differs from other OE approaches in several aspects. First, our ontology is informal: it contains concepts, their textual representations as terms, and semantic relatedness[3] as the only relation between concepts. To the best of our knowledge, we do not aware of approaches enriching ontology without at least taxonomic relations.

Second, we do not have domain-specific ontology that should be enriched explicitly; instead, we take general ontology that already has some domain-specific concepts (but we do not know explicitly which are) and add other domain-specific concepts to the whole ontology by analyzing collection of plain texts from the domain to be enriched.

This paper is organized as follows. Section 2 surveys related work in fields of ontology construction and enrichment; section 3 describes our approach in detail; in section 4 we present experimental results; the last section discusses the future work.

2. Related work

Ontology enrichment commonly means extracting new semantic relations [6, 32] or finding the appropriate place for domain-specific concepts in the existed taxonomy of the same domain [5, 25]. In this sense, our work is more related to general ontology learning, which has been widely surveyed [2, 3, 37], including our paper devoted to informal ontologies [1].

Drumond and Girardi [9] indentify so-called Ontology learning cake, see Figure 1.

---

[3] Function taking values from 0 (concepts have nothing in common) to 1 (concepts have the same semantics). Sometimes it is called ‘semantic similarity’ [21], but strictly, semantic similarity limits by ‘is-a’ relation [33].

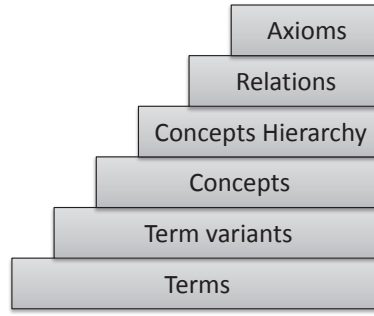


Figure 1. Ontology learning cake

The first layer, automatic terminology recognition (ATR), is the most researched one [8, 23, 27, 28, 29, 30]. In particular, we experimentally evaluated state-of-the-art approaches [11].

Term variants recognition has been studied a lot, too [7, 24, 27, 29], but most ATR works include only the simplest methods like stemming.

Regarding concepts formation, common approach is creating a concept for each term; it is based on observation that domain-specific term tends to have a single meaning [31].

Other layers correspond to formal ontologies only.

3. Ontology enrichment approach

This section describes our approach. Briefly, we take domain-specific texts and general ontology as input, then perform several sequential steps that are in keeping with Ontology learning cake, as a result we obtain domain-specific concepts (with terms and relations) that are missed in the input ontology. Each step is discussed below.

3.1. Preprocessing

At this step we apply the following methods to each input text:

1. Sentence detection
2. Tokenization
3. Part of speech tagging
4. Lemmatization
5. Word Sense Disambiguation

We used implementations from Texterra—our framework for text processing [35]. It uses, in turn, OpenNLP library[4] for the first 3 methods, heuristic algorithm based on morphologic properties of nouns for lemmatization, and Milne's algorithm [21] with another function of semantic relatedness [34] for WSD.

[4] http://opennlp.apache.org

3.2. Terminology recognition

This step takes all preprocessed texts and returns a set of domain-specific terms that are not contained in the input ontology. We adhere to the standard split of terminology extraction task [29]: collecting term candidates, computing features, and classifying term candidates into terms and not terms.

As term candidates we extract all uni-, bi-, and trigrams that occur at least 2 times and satisfy the following part of speech patterns: (N), (N_N), (Adj_N), (N_N_N), (Adj_N_N), (N_Adj_N), where N is noun and Adj is adjective.

We implemented most of state-of-the-art features, namely: CValue [13], MCValue [22], Lexical cohesion [28], Domain consensus [23], Domain relevance [27], Relevance [29], Weirdness [30], Frequency, Normalized frequency, TFIDF, Words count. As term classifier we use two approaches: Voting algorithm and supervised machine learning (ML) algorithm. The former combines features as follows:

$$V(t) = \sum_{i=1}^{n} \frac{1}{R(F_i(t))}$$

where t is a term candidate, n is a number of considered features, $R(F_i(t))$ is a rank of t among values of other term candidates considering feature F_i. Having ordered list of term candidates, one can take a top as most probable terms.

The second approach combines features in natural ML-way with Logistic regression as a particular algorithm.

Refer to our previous work [11] for details.

Since our aim is to enrich ontology, but not to construct it from scratch, we filter out terms already presented in the input ontology. For example, we enrich board game domain; term *board game* has an appropriate meaning in Wikipedia and thus should not be included into the final set of domain-specific concepts. However, one can suggest a counterexample: term *hand* in board games usually means *set of currently holding cards*, while Wikipedia has such term, but not such meaning. We describe the solution for this problem in the next subsection.

3.3. Concepts formation

There are two possible problems in transition from terms to concepts:

- synonymy—several terms have the same concept
- homonymy—several concepts have the same term

Currently we do not approach the former, partly because its effect is not so harmful, on conditions that relations for synonymic concepts are created correctly.

As for homonymy, we assume that domain-specific terminology is consistent [31] and does not contain homonyms inside the domain, i.e. we form a new concept for each newly extracted term. But we consider the case when term has a concept in general ontology and a domain-specific concept at once, see above for example with term

hand. To detect such terms, we use our approach [12]. Briefly, this method utilizes a binary logistic regression classifier based on the following features of the term:

- Relatedness to domain key concepts
- Domain relevance [27]
- Quotient of disambiguated concepts to the number of all occurrences of the term
- Quotient of disambiguated concepts to the number of possible existing concepts for the term.

3.4. Relations extraction

This section describes the way we extract relations for newly formed concepts, but firstly we discuss our ontology's organization [35]. As other systems based on Wikipedia knowledge [21], Texterra considers each article to be a concept and stores all incoming and outgoing links for an article. These links, or neighbor concepts, are used for semantic relatedness computing; in particular, Texterra uses Dice measure, that is a normalized number of common neighbors, but other measures are possible [34].

Thus, we seek at this step to extract concepts that are likely to be neighbors for each newly formed concept. If we look at distributional methods for synonyms detection [17], we can see that they are similar to the approaches for semantic relatedness computing. Indeed, the common algorithm is to collect contexts for input words, measure how similar they are, e.g. by Cosine or Dice, and extend the obtain value to the similarity between input words, on the assumption that "words that occur in the same contexts tend to have similar meanings" [15]. In this sense, neighbors represent context for concepts, therefore we take as neighbors for a concept those ones, which co-occur with the considered one not by chance. In order to find such concepts we adopt classical distributional methods for synonyms detection, particularly—measuring association with context [17].

More formally, for a newly formed concept we perform the following steps:

1. Collect neighbor candidates: for each term occurrence of each term of the input concept, find all term occurrences inside the specified window (15 occurrences to the left and to the right), and store their disambiguated concepts. As a result, we obtain a vector with concepts and their co-occurrence counts.
2. Transform each co-occurrence count into the more reliable value that shows randomness of co-occurrence: we use t-test measure with approximation of variance by sample mean [20].
3. Cut-off neighbor candidates by the predefined threshold, that is 2.0.

4. Evaluation

Approaches to evaluate ontology learning algorithm vary widely [37], because, first, ontology is not an end product; second, source corpora are usually huge and cannot be fully processed by human experts; and third, ontology construction process has

a complex structure. Indirect ontology evaluation by testing of applications that use the ontology (in-vivo testing) is not indicative in the case of bad results. It still requires direct ontology evaluation (in-vitro testing) in order to find bottleneck, if any. For these reasons we start from evaluation of each step separately.

As regards huge size of corpora, we propose the following: to markup manually only the small part of the corpus; to use the whole corpus in the prototype; to evaluate obtained results only for the marked up part. It allows having all available statistics in prototype, e.g. nonsparse counts of term occurrences, and, in the same time, obtaining reliable precision, recall, or other evaluation results.

4.1. Dataset overview

Evaluation based on a small proportion of marked up part imposes more severe requirements to the quality of the markup. At the same time, marking up each text by several people in order to average out mistakes is costly. Therefore we prepare Manifest[5]—detailed instructions for marking up text by the following annotations: terms, concepts (of the existing ontology), and domain-specificity (is the term of target domain, or another domain, or not domain-specific at all).

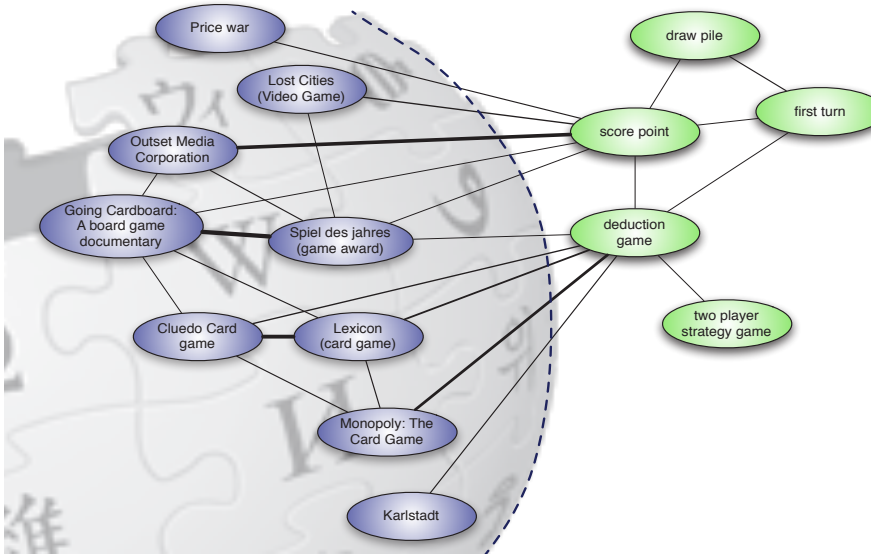


Figure 2. Examples of domain-specific concepts from marked up texts: left concepts are already presented in Wikipedia, right ones are new. Lines show semantic relatedness computed over links extracted by our tool. Thicker lines mean closer related concepts.

[5] Available on Russian at http://modis.ispras.ru/FTPContent/astrakhantsev/Manifest.pdf

We have chosen board game domain, since it is rather specific to be not fully presented in Wikipedia and is rather common to not require domain experts. We downloaded 1300 texts[6]—mostly user reviews and game specifications. 35 texts have been marked up by 9 humans (without overlapping). There have been found 1244 terms total, including 527 domain-specific ones, 246 domain-specific terms are missed in Wikipedia. Examples are shown in Figure 2.

4.2. Terminology recognition

We evaluated 2 scenarios:

1. Term recognition—compare terms without regard to their presence in Wikipedia
2. Term enrichment—compare only terms not presented in Wikipedia

We used 3 standard metrics here: average precision, precision, and recall (note that it is actually recall on candidates, i.e. fraction of term candidates that are correctly classified as terms; recall of our method for candidates collection is 71% for term recognition and 60% for term enrichment). We took top 500 of ranked candidates as terms, since it is the approximate count of domain-specific terms in marked up texts. For ML algorithm we used 2-fold cross-validation.

For each scenario we performed feature selection by exhaustive search. In case of several equally evaluated feature sets, we kept the smallest one. Results are shown in Table 1.

Table 1. Best feature subsets found by exhaustive search for Terminology recognition

		Best features subset	
		Precision, Recall	**Average Precision**
Term recognition	Voting	CValue, Relevance	MCValue, Relevance, TFIDF
	ML	Lexical cohesion, Words count, Cvalue, Relevance	Words count, CValue, Relevance
Term enrichment	Voting	CValue, Words count, TFIDF	MCValue, TFIDF
	ML	Lexical cohesion, TFIDF, Domain relevance, Relevance	Relevance, MCValue, TFIDF

Surprisingly, feature sets for Precision and Recall are the same, and this rule remains valid for all testing setups, that is why there is just one column for 2 metrics in Table 1. Another counter-intuitive observation is that ML algorithm does not select 'Words count' feature for Term enrichment.

Results for the best feature sets are presented in Table 2.

[6] From http://boardgamegeek.com

Table 2. Evaluation results for terminology recognition; the best feature subset was taken for each metric

		Precision	Recall	Average Precision
Term recognition	Voting	31.4	41.8	43.3
	ML	28.6	74.9	45.6
Term enrichment	Voting	18.2	61.9	34.1
	ML	13.8	92.0	33.0

As we can see, although implemented approach corresponds to state-of-the-art of automatic terminology recognition (and the same system shows much better results for other datasets, e.g. GENIA [11]), current results are too low to be applicable in practice. Nevertheless, based on high recall, we suppose semi-automatic methods to be promising for this task. In the simplest form, we can provide the expert with recognized terms along with contexts of occurrences and ask him/her to mark each one as domain-specific or not; tracking of user decisions could further improve productivity.

4.3. Concepts formation

To evaluate the approach for homonymy detection, we manually marked up 75 specific terms of the board games domain: 43 terms with existing concepts and 32 with new ones. We performed 4-fold cross-validation.

The results are presented in Table 3. Baseline means the method based on WSD confidence [10].

Table 3. Results for homonymy detection

	Precision	Recall	F-measure
Baseline	63%	67%	65%
Our approach	74%	83%	78%

As we can see, our approach significantly outperforms the baseline method.

4.4. Relation extraction

Unlike the other steps, extraction of semantic relatedness cannot be evaluated by direct comparison with manually prepared gold standard. However, as gold standard we can use domain-specific concepts that are already contained in Wikipedia. Normally we filter out terms of such concepts during the first step; for this evaluation scenario, we keep those automatically recognized terms, each of which has only one meaning in Wikipedia and this meaning is domain-specific.

Since we use extracted neighbors only for semantic relatedness computation, we compare this function for these two sets: neighbors that we extracted for newly

formed concepts (OE set) and neighbors that Wikipedia stores for the corresponding concepts. Moreover, we are interested in relative values of semantic relatedness; in particular, WSD uses this function to compare concepts by their relatedness to context. Therefore we evaluate the ranking of semantic relatedness function based on two neighbor sets. To be exact, we use mean average precision metric (MAP): having set of concepts, we choose one and rank others by their semantic relatedness to the chosen one; then we compute average precision for two lists obtained by both semantic relatedness; finally, we repeat the whole procedure for each concept in the input set and average these values.

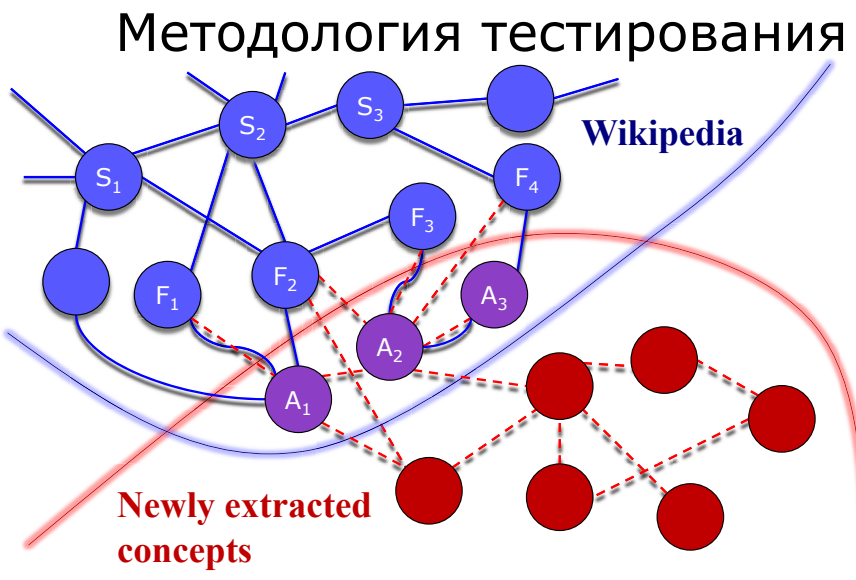


Figure 3. Sets of neighbors for evaluation of relation extraction algorithm. Blue solid lines mean links in Wikipedia; red dashed lines mean neighbors extracted by our system.

Since Dice measure is based on common neighbors, we compute MAP for 3 sets (see Figure 3):

- A—extracted concepts that are in Wikipedia
- F—first neighbors of A in OE set
- S—second neighbors of A in OE set (we use subset of 2000 concepts due to performance issue)

Results are shown in Figures 4–6.

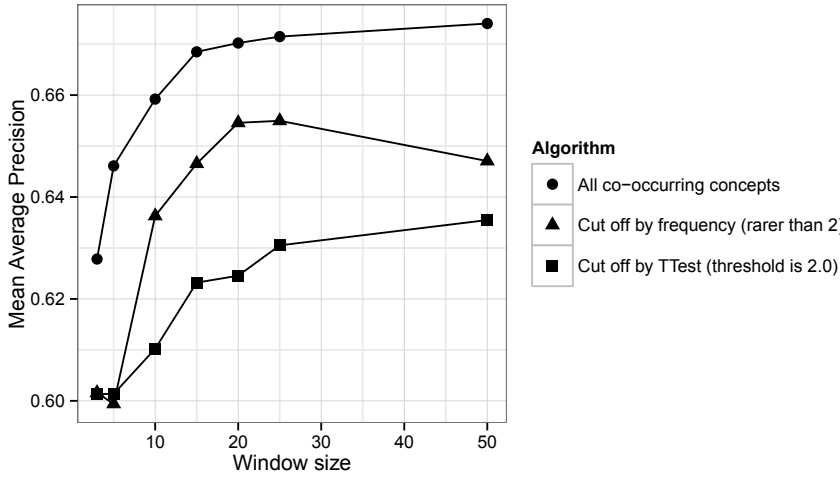


Figure 4. Mean Average Precision over A (extracted concepts that are in Wikipedia)

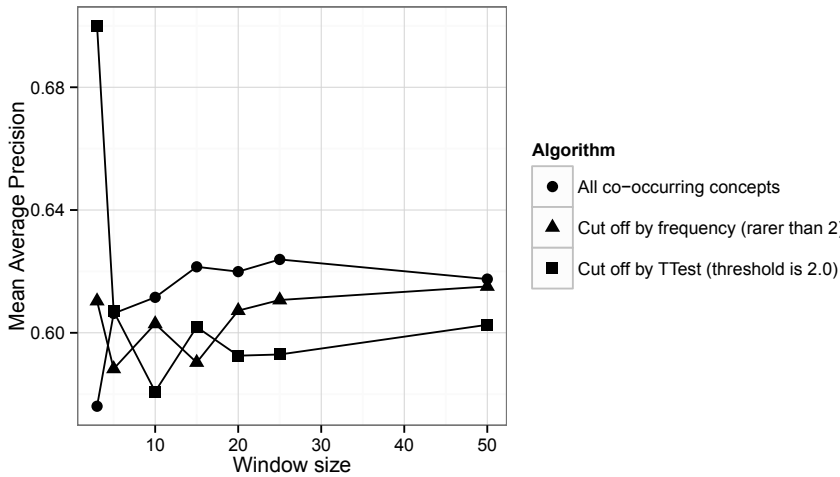


Figure 5. Mean Average Precision over F (first neighbors of A)

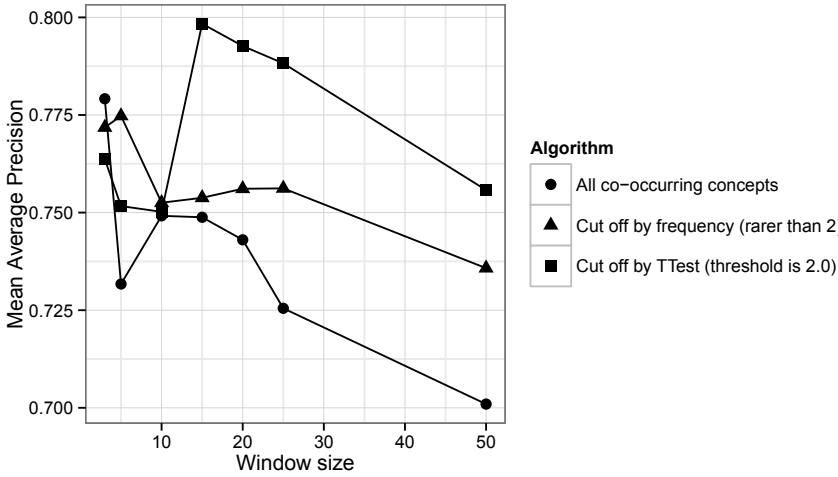


Figure 6. Mean Average Precision over S (second neighbors of A)

Note that we compute average precision between two lists of the same concepts (ordered by different algorithms), so in worst case—our semantic relatedness has nothing in common with Wikipedia's one—it would be 0.5. Our results are not so higher than 0.5, we believe it happens because there are few occurrences for most concepts, so there are few co-occurring concepts, so we cover less context than Wikipedia. It also explains why taking a huge amount of concepts as neighbors does not degrade results and why results for A improve with increasing window size: extracted neighbors serve mostly to connect to Wikipedia and to enable its set to participate in semantic relatedness computing.

5. Conclusions and Future work

This work addresses the problem of ontology incompleteness, in particular for wikification systems. We formulated the task as ontology construction from domain-specific texts and broke it into the steps according to Ontology learning cake. We used state-of-the-art approach for terminology extraction; our own method—for concepts formation; classic methods from the related field—for relations extraction. Another contribution is the methodology for dataset preparation and its usage for each step evaluation.

The main drawback of this work is low precision of terminology extraction. We plan to experiment with semi-automatic approaches, since recall is rather acceptable. In addition, if we ask the user to firstly provide domain-specific concepts that are already presented in Wikipedia, the modified task would be similar to *entity set expansion*, which has a lot of promising approaches [18][8]. Besides, currently we do not use existing ontology, although it contains necessary background knowledge; we are going to implement bootstrapping approach: to correct term recognition mistakes on the basis of extracted relations.

Also we plan to extend current dataset and to test on other domains, including cross-domain evaluation.

References

1. *Astrakhantsev, N. A., Turdakov, D. Y.* (2013). Automatic construction and enrichment of informal ontologies: A survey. Programming and Computer Software, 39(1), 34–42.
2. *Biemann, C.* (2005). Ontology Learning from Text: A Survey of Methods. In LDV forum (Vol. 20, No. 2, pp. 75–93).
3. *Buitelaar, P., Cimiano, P.* (Eds.). (2008). Ontology learning and population: bridging the gap between text and knowledge (Vol. 167). Ios Press.
4. *Cheng, X., Roth, D.* (2013). Relational Inference for Wikification. Urbana, 51, 61801.
5. *Chifu, E. T., Le Ia, I. A.* (2008). Text-based ontology enrichment using hierarchical self-organizing maps.
6. *Cimiano, P., Hotho, A., Staab, S.* (2005). Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. J. Artif. Intell. Res. (JAIR), 24, 305–339.
7. *Daille, B., Habert, B., Jacquemin, C., Royauté, J.* (1996). Empirical observation of term variations and principles for their description. Terminology, 3(2), 197–257.
8. *Dalvi, B., Callan, J., & Cohen, W.* (2011). Entity list completion using set expansion techniques. Carnegie-Mellon Univ Pittsburgh Pa Language Technologies Inst.
9. *Drumond, L., Girardi, R.* (2008). A Survey of Ontology Learning Procedures. WONTO, 427.
10. *Erk, K.* (2006). Unknown word sense detection as outlier detection. In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (pp. 128–135). Association for Computational Linguistics.
11. *Fedorenko, D. G., Astrakhantsev, N. A., Turdakov, D. Y.* (2013). Automatic recognition of domain-specific terms: an experimental evaluation. SYRCoDIS (pp. 15–23).
12. *Fedorenko, D. G., Astrakhantsev, N. A.* (2013). Automatic Extraction of New Concepts from Domain-Specific Terms [Izvlechenie novykh kontseptov predmetnospetsifichnykh terminov]. Proceedings of the Institute for System Programming of RAS, volume 25 (pp. 167–178).
13. *Frantzi, K. T., Ananiadou, S.* (1996). Extracting nested collocations. In Proceedings of the 16th conference on Computational linguistics-Volume 1 (pp. 41–46). Association for Computational Linguistics.
14. *Grineva, M., Grinev, M., Lizorkin, D., Boldakov, A., Turdakov, D., Sysoev, A., Kiyko, A.* (2011, March). Blognoon: exploring a topic in the blogosphere. In Proceedings of the 20th international conference companion on World wide web (pp. 213–216). ACM.
15. *Harris, Z. S.* (1954). Distributional structure. Word.
16. *Jimeno-Yepes, A., Berlanga-Llavori, R., Rebholz-Schuhmann, D.* (2010). Ontology refinement for improved information retrieval. Information Processing & Management, 46(4), pp. 426–435.
17. *Jurafsky, D., James, H.* (2000). Speech and language processing an introduction to natural language processing, computational linguistics, and speech.

18. *Li, X. L., Zhang, L., Liu, B., & Ng, S. K.* (2010, July). Distributional similarity vs. PU learning for entity set expansion. In Proceedings of the ACL 2010 Conference Short Papers (pp. 359–364). Association for Computational Linguistics.
19. *Lin, T., Etzioni, O.* (2012). No noun phrase left behind: detecting and typing unlinkable entities. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 893–903). Association for Computational Linguistics.
20. *Manning, C. D.* (1999). Foundations of statistical natural language processing. H. Schütze (Ed.). MIT press.
21. *Milne, D., Witten, I. H.* (2008). Learning to link with wikipedia. In Proceedings of the 17th ACM conference on Information and knowledge management (pp. 509–518). ACM
22. *Nakagawa, H., Mori, T.* (2002). A simple but powerful automatic term extraction method. In COLING-02 on COMPUTERM 2002: second international workshop on computational terminology-Volume 14 (pp. 1–7). Association for Computational Linguistics.
23. *Navigli, R., Velardi, P.* (2002). Semantic interpretation of terminological strings. In Proc. 6th Int'l Conf. Terminology and Knowledge Eng (pp. 95–100).
24. *Nenadié, G., Ananiadou, S., McNaught, J.* (2004). Enhancing automatic term recognition through recognition of variation. In Proceedings of the 20th international conference on Computational Linguistics (p. 604). Association for Computational Linguistics.
25. *Neshati, M., Alijamaat, A., Abolhassani, H., Rahimi, A., Hoseini, M.* (2007, November). Taxonomy learning using compound similarity measure. In Web Intelligence, IEEE/WIC/ACM International Conference on (pp. 487–490). IEEE.
26. *Noy, N. F., Klein, M.* (2004). Ontology evolution: Not the same as schema evolution. Knowledge and information systems, 6(4), 428–440.
27. *Park, Y., Byrd, R. J., Boguraev, B. K.* (2002). Automatic glossary extraction: beyond terminology identification. In Proceedings of the 19th international conference on Computational linguistics-Volume 1 (pp. 1–7). Association for Computational Linguistics.
28. *Patry, A., Langlais, P.* (2005). Corpus-based terminology extraction. In Proceedings of the 7th International Conference on Terminology and Knowledge Engineering (pp. 313–321).
29. *Pazienza, M. T., Pennacchiotti, M., Zanzotto, F. M.* (2005). Terminology extraction: an analysis of linguistic and statistical approaches. In Knowledge Mining (pp. 255–279). Springer Berlin Heidelberg.
30. *Peñas, A., Verdejo, F., Gonzalo, J.* (2001). Corpus-based terminology extraction applied to information access. In Proceedings of Corpus Linguistics (Vol. 2001).
31. *Slozhenikina, J. V.* The Term: Real as Life (Why Term Can and Should Have Variants), [Termin: zhivoi kak zhizn' (pochemu termin mozhet i dolzhen imet' varianty)] "Znanie. Ponimanie. Umenie", 2010, vol. 5
32. *Schutz, A., Buitelaar, P.* (2005). Relext: A tool for relation extraction from text in ontology extension. In The Semantic Web–ISWC 2005 (pp. 593–606). Springer Berlin Heidelberg.

33. *Strube, M., Ponzetto, S. P.* (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In AAAI (Vol. 6, pp. 1419–1424).
34. *Turdakov, D., Velikhov, P.* (2008). Semantic relatedness metric for wikipedia concepts based on link analysis and its application to word sense disambiguation. SYRCoDIS, In CEUR Workshop Proceedings, vol. 355.
35. *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, volume 26, Issue 1, pp. 421–438
36. *Unger, C., Cimiano, P.* (2011). Pythia: Compositional meaning construction for ontology-based question answering on the Semantic Web. In Natural Language Processing and Information Systems (pp. 153–160). Springer Berlin Heidelberg.
37. *Wong, W., Liu, W., Bennamoun, M.* (2012). Ontology learning from text: A look back and into the future. ACM Computing Surveys (CSUR), 44(4), 20.

АКТИВНОСТЬ УЧАСТНИКА КОММУНИКАЦИИ: МЕТОДЫ ЛИНГВИСТИЧЕСКОГО АНАЛИЗА

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

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

ACTIVITY OF PARTICIPANTS IN A CONVERSATION: METHODS OF LINGUISTIC ANALYSIS

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

Vinogradov Russian Language Institute of the Russian Academy of Sciences, Russia

The paper deals with the phenomenon of activity of dialogue participants. Analysis of participants' activity in a conversation is of great importance for theoretical linguistics as well as for applied linguistics. In forensic linguistics, analysis of activity can be used as an objective parameter for the qualification of real communicative goals of participants. The paper introduces three major methods of analysis of the phenomenon discussed: 1) the method of **communicative activity**, i.e. the amount of illocutionary independent speech acts of a participant in a dialogue or in its relevant part; 2) the method of **thematic activity**, the analysis of which enables the detection of exactly which participants independently introduces the main themes in a conversation; 3) the method of **quantitative activity**, based on calculating the amount of words associated with a specific theme in a conversation. We discuss the different types of correlation between the three methods.

Постановка проблемы

Речевое поведение в диалоге можно анализировать по различным параметрам, определяемым как теоретическими рамками исследования, так и разнообразными практическими задачами. В рамках социологической проблематики

важно, например, выявить возможную связь между социальным статусом участника и его участием в диалоге. Так, в рамках военного дискурса приказ может исходить только от лица, занимающего более высокую ступень в военной иерархии. В некоторых культурных традициях наблюдаются ограничения в формах участия в коммуникации женщин и детей. Этот аспект общения исследуется в рамках этнометодологии [Гарфинкель 2007; Garfinkel 1967] и анализа разговора как одного из методов этнометодологии и одновременно автономного направления исследований обыденного общения (conversation analysis). Психологический анализ сосредотачивается на психологических мотивациях решений, принимаемых участниками, на психологическом обосновании целеполагания. В дискурс-анализе обращается внимание на различные особенности взятия инициативы участником — прерывание собеседника, навязывание собственной темы, игнорирование вопросов, советов и требований собеседника и пр. Некоторые аспекты активности участников диалога обсуждались в социолингвистике [Таннен 2012].

В собственно лингвистической традиции исследования общения активность участника коммуникации не рассматривалась как релевантный языковой параметр и лингвистические модели не предусматривали такое направление анализа речевого поведения говорящих. Между тем в приложениях лингвистики — в частности в лингвистической экспертизе текста — данный параметр весьма значим для выявления истинных коммуникативных целей участников. Наиболее существенна оценка активности участников диалога в экспертизах по делам о взятках (в том числе о провокации взятки) и вымогательстве (ст. 290, 291, 304; 163 УК РФ). Действительно, участники диалогов о взятках — как взяткодатель, так и взяточник — во многих случаях прекрасно осведомлены о возможных санкциях за дачу взятки, за получение взятки и за провокацию взятки. Тем самым, прямое выражение соответствующих коммуникативных намерений участниками диалогов о взятках всячески избегается. Исключение составляют ситуации «бытовой взятки» — дополнительное (не предусмотренное тарифной сеткой и соответствующим списком дополнительных услуг) вознаграждение сантехника, электрика, работника социальной сферы, сотрудника автоинспекции и под. В этих случаях маскировка коммуникативных целей участников в подавляющем большинстве случаев отсутствует.

Аналогичная ситуация в делах о вымогательстве: вымогатель, как правило, осознает противоправность собственного поведения. Кроме того, часто он подозревает, что его реплики записываются. В такой ситуации маскировка истинных намерений необходима. Однако активность поведения участника в диалоге — особенно по какой-то конкретной теме — в совокупности со специфическими языковыми маркерами сокрытия коммуникативного намерения дает возможность эксперту прийти к выводу о том, что тот или иной участник пытается уйти от прямого выражения своих коммуникативных намерений и избегнуть санкций по соответствующей статье УК РФ. Следует отметить, что сам по себе анализ активности участника в рамках той или иной темы общения недостаточен и должен быть дополнен исследованием эксплицитной (пропозициональной) составляющей диалога, а также изучением скрытых, косвенных и неявных способов передачи семантики. Это не входит в задачи настоящей работы, поскольку требует особого рассмотрения.

В настоящей работе описываются три метода анализа активности участников диалога (мекоторые позволяют количественно — и тем самым, достаточно объективно — квалифицировать речевое поведение участников коммуникативного взаимодействия. Предлагаемые вности допускают алгоритмизацию и компьютерную реализацию в соответствующих программах автоматической обработки текста.

Метод анализа коммуникативной активности

Интенсивность участия в общении можно определять количеством реплик (речевых актов), которые независимы с коммуникативной точки зрения от других реплик, но которые побуждают реагировать собеседника — отвечать, давать оценку, соглашаться или не соглашаться и т. д. В теории речевых актов такие случаи описываются в терминах «иллокутивно вынуждающих» и «иллокутивно вынуждаемых» речевых актов. Иллокутивно вынуждающий речевой акт не зависит с точки зрения коммуникативного намерения ни от каких других речевых актов в диалоге, а иллокутивно вынуждаемый речевой акт зависит по коммуникативной функции от других речевых актов (см. подробнее по этому поводу: [Баранов, Крейдлин 1992 а, б]). Так, в последовательности вопрос — ответ, требование — речевое выражение подчинения или отказ, совет — согласие/отказ и т. д. первые (в парах) речевые акты (вопрос, требование, совет) являются иллокутивно вынуждающими, а вторые (ответ, речевое выражение подчинения, отказ, согласие) — иллокутивно вынуждаемыми.

Количество независимых реплик, коммуникативно вынуждающих реплики партнера по общению, определяет коммуникативную активность участника в обсуждении той или иной теме беседы. Этот показатель называется **параметром коммуникативной активности**.

Приведем пример реального диалога, представленного эксперту в рамках лингвистической экспертизы о взятке[1]:

Ж (*входя в кабинет*) — Можно? Это вам, просили передать
по энергосбережению — сегодня была учеба по мероприятиям.
И я у Вас хотела встречный вопрос спросить. По поводу…
М — Кто это прислал?
Ж — У нас сегодня приходили предприниматели, и они вели обучение,
вот. А обучение проводили и сказали отдать главе управы. Что бы
он ознакомился. Второй вопрос, мы выводим из схемы хозтовары?
М — Ну а зачем это нужно? Оставим как есть.
Ж — И третий вопрос. По поводу павильонов около железной дороги.
Значит, Поляков. Я встречалась с Поляковым, вот, как бы
он спрашивает, можно ли сумму поменьше сделать

[1] Обмен репликами в методических целях несколько упрощен и нормирован, опущена обсценная лексика.

[*Мужчина отрицательно качает головой, поднимает правую руку вверх с выпрямленным указательным пальцем*.]

Ж — Это что это?
М — Ну да. Это так называется.
Ж — Понятно. Один миллион?
М — Да
Ж — Понятно. И потом как бы, гарантии какие?
М — Трогать никто не будет.
Ж — Никто не будет трогать. Понятно. И еще такой вопрос. По поводу оплаты…
М — Прическу поменяла?
Ж — Да не поменяла прическу. Я просто…
М — Подстриглась?
Ж — Нет, я помыла голову и уложилась феном. По поводу оплаты… Он спрашивает, возможно ли отсрочка на немного?
М — Нет.
Ж — Нет. Хорошо… Безналом можно?
М — Все равно…
Ж — Рублями? В долларах?
М — Безразлично…
Ж — Ясно, хорошо.

В рассматриваемом фрагменте диалога коммуникативно независимыми являются следующие реплики участницы Ж:

Можно?

Это вам, просили передать по энергосбережению — сегодня была учеба по мероприятиям.

И я у Вас хотела встречный вопрос спросить.

Второй вопрос, мы выводим из схемы хозтовары?

Я встречалась с Поляковым, вот, как бы он спрашивает, можно ли сумму поменьше сделать

Один миллион?

И потом как бы, гарантии какие?

И еще такой вопрос. По поводу оплаты…

Он спрашивает, возможно ли отсрочка на немного?

Безналом можно?

Рублями? В долларах?

Итого выделяется одиннадцать коммуникативно независимых реплик участницы Ж.

У участника М в исследуемом фрагменте диалога обнаруживаются следующие коммуникативно независимые реплики:

Кто это прислал?

Прическу поменяла?

Подстриглась?

Общее число коммуникативно независимых реплик участника М — три.

Остальные реплики участников либо коммуникативно зависят от других реплик другого участника, либо являются реакциями на его невербальные действия, либо иллокутивно самовынуждаются тем же участником (данные реплики несущественны с точки зрения рассматриваемого метода и, соответственно, параметра активности).

Таким образом, по параметру коммуникативной активности участница Ж существенно превосходит участника М: одиннадцать реплик против трех. Отметим, правда, что четыре иллокутивно независимых реплики участницы Ж не относятся к теме денег:

Можно?

Это вам, просили передать по энергосбережению — сегодня была учеба по мероприятиям.

И я у Вас хотела встречный вопрос спросить.

Второй вопрос, мы выводим из схемы хозтовары?

Более важный аспект изучения активности — это коммуникативная активность участников при обсуждении конкретной темы, релевантной для экспертизы. В рассматриваемом случае важной оказывается тема «Поляков-выплата денег», которая представлена в следующем обмене репликами участников:

Ж — И третий вопрос. По поводу павильонов около железной дороги. Значит, Поляков. Я встречалась с Поляковым, вот, как бы он спрашивает, можно ли сумму поменьше сделать

[*Мужчина отрицательно качает головой, поднимает правую руку вверх с выпрямленным указательным пальцем*.]

Ж — Это что это?
М — Ну да. Это так называется.
Ж — Понятно. Один миллион?
М — Да
Ж — Понятно. И потом как бы, гарантии какие?
М — Трогать никто не будет.
Ж — Никто не будет трогать. Понятно. И еще такой вопрос. По поводу оплаты…
<…>
Ж — <…> По поводу оплаты… Он спрашивает, возможно ли отсрочка на немного?
М — Нет.
Ж — Нет. Хорошо… Безналом можно?
М — Все равно…
Ж — Рублями? В долларах?
М — Безразлично…
Ж — Ясно, хорошо.

В этом фрагменте участница Ж вводит в диалог пять иллокутивно независимых реплик:

Ж — <…> Я встречалась с Поляковым, вот, как бы он спрашивает, можно ли сумму поменьше сделать
Ж — <…> И потом как бы, гарантии какие?
Ж — <…> По поводу оплаты… <…> По поводу оплаты… Он спрашивает, возможно ли отсрочка на немного?
Ж — <…> Безналом можно?
Ж — Рублями? В долларах?

В речевом поведении участника М иллокутивно независимые реплики при обсуждении темы «Поляков-выплата денег» отсутствуют. Таким образом, участница Ж по данному параметру и применительно к теме денег существенно активнее участника М: пять против ноля реплик.

Метод анализа содержательной активности

Существенным для характеристики активности в диалоге оказывается также определение того участника, который коммуникативно независимо вводит основные темы беседы. Этот показатель называется **параметром содержательной активности**. Данный параметр указывает на то, кто определяет общее содержание коммуникации. Действительно, участник, формирующий набор тем

беседы, определяет ее содержание. Основные темы в рассматриваемом фрагменте диалога — «Обучение», «Хозтовары», «Поляков-выплата денег», «Прическа». Первые три темы содержательного характера вводит в диалог участница Ж:

> Это вам, просили передать по энергосбережению — сегодня была учеба по мероприятиям.
>
> Второй вопрос, мы выводим из схемы хозтовары?
>
> И третий вопрос. По поводу павильонов около железной дороги. Значит, Поляков. Я встречалась с Поляковым, вот, как бы он спрашивает, можно ли сумму поменьше сделать

Участник М вводит только одну тему — «Прическа»:

> Прическу поменяла?

Таким образом, в рассматриваемом случае основные темы диалога вводятся участницей Ж. И по этому параметру она оказывается более активным участником по сравнению с участником М. Следует отметить, что тема «Прическа», которую вводит участник М, не представляет интереса с точки зрения проводимой экспертизы, поскольку она никак не связана с передачей денег, вознаграждением и т. д.

Метод анализа количественной активности

Активность участника проявляется также в количестве словоформ, использованных в обсуждении темы беседы (количество словоупотреблений). Эта характеристика называется параметром **количественной активности**. Параметр количественной активности осмысленно рассматривать только применительно к конкретной теме, которая обсуждается в диалоге и которая представляет интерес для экспертного исследования. В данном случае речь идет о теме «Поляков-выплата денег». Реплики участницы Ж, относящиеся к этой теме таковы:

> Ж — И третий вопрос. По поводу павильонов около железной дороги. Значит, Поляков. Я встречалась с Поляковым, вот, как бы он спрашивает, можно ли сумму поменьше сделать
> Ж — Это что это?
> Ж — Понятно. Один миллион?
> Ж — Понятно. И потом как бы, гарантии какие?
> Ж — Никто не будет трогать. Понятно. И еще такой вопрос. По поводу оплаты…
> Ж — <…> По поводу оплаты… Он спрашивает, возможно ли отсрочка на немного?

Ж — Нет. Хорошо… Безналом можно?
Ж — Рублями? В долларах?
Ж — Ясно, хорошо.

Приведенные реплики участницы Ж включают 68 словоформ.

Реплики участника М при обсуждении темы денег содержат существенно меньше словоформ — 15:

М — Ну да. Это так называется.
М — Да
М — Трогать никто не будет.
М — Нет.
М — Все равно…
М — Безразлично…

Таким образом, и поданному параметру количественной активности участница Ж существенно более активен, чем участник М.

Отмечу, что метод определения количественной активности похож по своим методологическим основаниям на контент-анализ в его классическом варианте (см., например, [Krippendorff 1980; Баранов 2000]), однако об оценке репрезентативности в случае параметра количественной активности говорить не приходится из-за незначительного объема выборки. Впрочем, и сам объект исследования в данном случае ограничен по объему.

Проблема интерпретации

Во всех случаях исследований речевого материала, связанных с разнообразными подсчетами, возникает проблема интерпретации полученных статистических данных. В лингвистической экспертизе квалификация активности участника по сути представляет собой промежуточный результат, предварительные данные, которые подлежат дальнейшей интерпретации либо самим экспертом, либо инстанцией, назначившей экспертизу (дознавателем, следователем, судьей и т. д.). Понятно, что активность участника диалога по всем рассмотренным параметрам указывает на то, что именно он инициирует дискурс о взятке, контролирует его и пытается воздействовать на собеседника в ту или иную сторону. То есть, если активность по всем параметрам касается взяточника, то он провоцирует собеседника на то, чтобы тот дал взятку — «кошмарит» его. С другой стороны, если активность по всем параметрам распространяется на дающего взятку или посредника (как в рассматриваемом случае), то весьма вероятна провокация взятки — вынуждение партнера по коммуникации взять деньги. Понятно также, что выявление точного содержания воздействия должно основываться на внимательном изучении пропозициональной семантики соответствующих реплик.

В рассматриваемом случае чрезмерная активность предполагаемого посредника (участницу, обозначенную буквой «Ж») подозрительна. Действительно, участница Ж сама поднимает вопрос о деньгах: *Я встречалась с Поляковым, вот, как бы он спрашивает, можно ли сумму поменьше сделать*, вынуждая адресата — участника М — указать конкретную сумму, которую он хочет получить за содействие. Участник М, не без оснований полагая, что его могут записывать, сразу не указывает сумму и делает неопределенный жест рукой. В официальной фоноскопии этот жест описан так: *Мужчина отрицательно качает головой, поднимает правую руку вверх с выпрямленным указательным пальцем*. Это описание некорректно. Замедленное воспроизведение видеоролика с записью беседы показывает, что участник М, совершая жест, отклоняет палец назад в сторону правого плеча, затем совершает пальцем полукруг в сторону и лишь затем фиксирует палец перед собой на уровне чуть выше глаз. Общепринятый жест, обозначающий «один», не содержит этих дополнительных динамических невербальных компонентов. Тем самым, жест, совершаемый участником М, не является конвенциональным, то есть общепринятым, и ему не соответствует общепринятой семантики. Наличие этих дополнительных компонентов приводит к тому, что участница Ж не понимает жеста участника М и спрашивает его: *это что это?* Наконец, она сама берет инициативу на себя: *Понятно. Один миллион?* Участник М соглашается с интерпретацией: *Да*. Участница Ж постоянно возвращается к теме денег, причем неспровоцированно, по своей собственной инициативе. Так, она спрашивает о возможности отсрочки (*возможно ли отсрочка на немного?*), о гарантиях помощи со стороны участника М (*И потом как бы, гарантии какие?*), о вариантах оплаты (*Безналом можно?*). Таким образом, исследование факторов активности в диалоге и пропозициональной составляющей показывает, что участница Ж провоцирует участника М на получение взятки, однако участник М отнюдь не возражает против этого: в других диалогах он в неявном виде указывает на санкции, которые последуют в случае, если деньги не будут ему переданы (*шлепнем сейчас их хорошенько; снесем просто их да и все; бульдозер приедет и все уберут и все*). Поскольку видеозапись ведет участница Ж, то провокация с ее стороны имеет вынужденный характер — это попытка доказательно зафиксировать коммуникативные намерения участник М.

Более сложная интерпретация требуется в тех случаях, когда возникает конкуренция активности участников по указанным параметрам. Так, при расхождении коммуникативной активности и содержательной активности результирующий вывод неочевиден. Действительно, и взяточник и взяткодатель могут получить соответствующую квалификацию по данным параметрам. Определенного вывода о провокации взятки или о самой взятке по таким данным не сделаешь. Для этого опять-таки необходим анализ пропозициональной семантики. В то же время параметр количественной активности в рассматриваемом типе диалогов оказывается более слабым по сравнению с параметрами коммуникативной и содержательной активности. При конкуренции параметров его разумно использовать как вспомогательный, позволяющий принять обоснованное решение после изучения факторов коммуникативной

и содержательной активности, а также после исследования пропозициональной семантики реплик участников.

Литература

1. *Баранов А. Н.* (2000), Введение в прикладную лингвистику. М.: УРСС.
2. *Баранов А. Н., Крейдлин Г. Е.* (1992а), Языковое взаимодействие в диалоге и понятие иллокутивного вынуждения // Вопросы языкознания, № 2
3. *Баранов А. Н., Крейдлин Г. Е.* (1992б) Структура диалогического текста: лексические показатели минимальных диалогов // Вопросы языкознания, 1992б.
4. *Гарфинкель Г.* (2007) Исследования по этнометодологии. СПб.: Питер.
5. *Таннен Д.* (2012) Коммуникативный стиль нью-йоркских евреев // Социолингвистика и социология языка. Хрестоматия. СПб.
6. *Garfinkel H.* Studies in Ethnomethodology. Englewood Cliffs; N. Y., 1967.
7. *Krippendorff K.* (1980), Content analysis. An introduction to its methodology. Lnd.

References

1. *Baranov A. N.* (2000), Introduction into applied linguistics [Vvedenije v prikladnuju lingvistiku], URSS, Moscow.
2. *Baranov A. N., Krejdlin G. E.* (1992a), Language interaction in a dialogue and the notion of illocutionary constraining [Jazykovoje vzaimodejstvije v dialoge i ponyatije illokutivnogo vynuzhdenija], Questions of linguistics [Voprosy jazykoznanija], № 2.
3. *Baranov A. N., Krejdlin G. E.* (1992b), Structure of a text of a dialogue: lexical markers of minimal dialogues [Struktura dialogicheskogo teksta: leksicheskije pokazateli minimalnyh dialogov], Questions of linguistics [Voprosy jazykoznanija], № 3.
4. *Garfinkel H.* (2007) Studies in Ethnomethodology [Issledovanija po etnometodologii]. SPb: Piter.
5. *Garfinkel H.* (1967) Studies in Ethnomethodology. Englewood Cliffs; N. Y.
6. *Krippendorff K.* (1980), Content analysis. An introduction to its methodology. Lnd.
7. *Tannen D.* (2012) New York Jewish Conversational Style [Kommunikativnyj stil' nju-jorkskih jevreev], Sociolinguistics and sociology of language [Sotiolingvistika i sociologija jazyka]. SPb.

MULTIMODAL AND CROSS-MODAL DISTRIBUTIONAL SEMANTICS: TOWARDS COMMON SEMANTIC SPACE FOR WORDS AND THINGS

Baroni M. (marco.baroni@unitn.it)

Center for Mind/Brain Sciences, University of Trento, Italy

Distributional semantic models (DSMs) capture various aspects of word meaning with vectors summarizing their patterns of co-occurrence in large text corpora, under the assumption that the contexts in which words occur are good cues of what they mean. DSMs have been very successful empirically, and they have been used to model increasingly sophisticated linguistic and cognitive phenomena.

However, current DSMs account for linguistic meaning entirely in terms of linguistic signs (the "meaning" of a word is a summary of the linguistic contexts in which the word occurs). This leads to two big conceptual problems: lack of grounding and lack of reference. Concerning the former, cognitive scientists have accumulated plenty of evidence that, for human beings, meaning is strongly embodied in the sensory-motor system, so a semantic theory that completely dissociates meaning from perception and action is, a priori, a rather implausible model of how humans work — a fact that has also empirical consequences in the surprisingly bad performance of DSMs on simple tasks requiring perceptual information. Lack of reference is perhaps an even more serious problem. A theory that has no way to connect the semantic representation of a linguistic expression to states of the world is clearly missing something fundamental about language, as it has no way to explain how we can talk about things!

Interestingly, in the last decade, it has become common in computer vision to represent images through vectors recording the distribution of automatically extracted discrete visual features in them — a representation that is very similar to the one that DSMs assume for words. This suggests that we might be able to free DSMs from their textual cage by establishing a connection with the visual world by means of such vector-based image-representation techniques.

In my talk, after a brief general introduction to distributional semantics, I will discuss experiments we carried out in the last few years in which we tackle the grounding problem (DSMs with richer multimodal semantic representations that combine linguistic and visual features), and recent work in which we started dealing with the reference issue (how to map images and linguistic expressions across modalities to a common space, in order to link language to the world out there). The case studies I will present include simulating human semantic similarity judgments, predicting the color of objects, modeling brain data and learning names and verbally-expressed attributes of objects present in pictures from indirect evidence.

ДИФФЕРЕНЦИАЛЬНАЯ КОРПУСНАЯ СТАТИСТИКА НА ОСНОВАНИИ НЕАВТОМАТИЧЕСКОЙ МЕТАТЕКСТОВОЙ РАЗМЕТКИ

Беликов В. И. (vibelikov@gmail.com)
РГГУ, Москва, Россия

Копылов Н. Ю. (Nikolay_Ko@abbyy.com)
РГГУ; ABBYY, Москва, Россия

Селегей В. П. (Vladimir_S@abbyy.com)
РГГУ; МФТИ; ABBYY, Москва, Россия

Шаров С. А. (s.sharoff@leeds.ac.uk)
РГГУ, Москва, Россия; University of Leeds, Великобритания

Статья основывается на исследовательских работах, проводящихся в рамках проекта создания Генерального Интернет-Корпуса Русского Языка (ГИКРЯ). В настоящее время одной из самых актуальных задач, ждущих своего решения в проекте, является автоматическая метатекстовая разметка. Тем не менее, в первую пробную версию корпуса включено большое количество материала, позволяющее проводить дифференциальный статистический анализ на большом объеме размеченных данных из разных сегментов интернета уже сейчас.

В настоящее время растет понимание того, что объем данных в ручных корпусах недостаточен для многих типов лингвистических исследований. При этом идея, что не только размер имеет значение, еще не стала настолько же популярной. В данной работе мы пытаемся показать критическую важность дифференциального анализа материала в корпусах-миллиардниках.

VARIATIONAL CORPUS STATISTICS USING AUTHOR PROFILES

Belikov V. (vibelikov@gmail.com)
RSUH, Moscow, Russia

Kopylov N. (Nikolay_Ko@abbyy.com)
RSUH, ABBYY, Moscow, Russia

Selegey V. (Vladimir_S@abbyy.com)
RSUH, ABBYY, Moscow, Russia

Sharoff S. (s.sharoff@leeds.ac.uk)
RSUH, Moscow, Russia; University of Leeds, UK

This paper is based on research carried out in the framework of our project on the General Internet Corpus of Russian (Geekrya) . The need to use large-scale corpora automatically collected from the Web was first recognized in computational linguistics. Recently, the lack of data in "manually-built" corpora led to recognition of the importance of Web-derived corpora in traditional linguistic research.

The principal difference of Geekrya from the two other large web corpora of Russian (RuWac and RuTenTen) is that the latter were produced by indiscriminate crawling of the Russian Internet, resulting in no metatext markup available for their data.

GEEKRYA is different since its contents is split into "segments" which we define as a compact set of webpages sharing a general communicative purpose expressed in text-rich content. We extracted information about the authors from their profiles when this was specified.

The total size of indexed Geekrya amounts to 12 billion words, the segments with known a priori metatext parameters are listed below (size given in millions of words).

Segment	Gender	Age	Region
blogs.mail.ru	164	81	113
livejournal.com	0	1,800	5,600
vk.com	2,000	1,600	1,600
news	0	0	0
magazines.russ.ru	258	0	0
forums (adw.ru)	163	0	0
Total:	2,585	3,481	7,313

The magazines.russ.ru segment, for example, contains all the texts from this resource (mostly published fiction and literary criticism). Author's gender has been extracted for 84.3% its texts, the size of the male subcorpus is—194 MW, the female one is 64 MW.

Information about the author' profiles within the individual segments helps in variational analysis. The paper lists several studies on the gender profiles of discourse words, collocations and idioms, as well as on the regional distribution, for example, comparing word uses in Siberia against the rest of the Russian-speaking world.

Мегакорпуса русскоязычного интернета

Необходимость привлечения к языковому анализу автоматически собранных корпусов была осознана сначала в компьютерной лингвистике. Но в последнее время нехватка данных в корпусах «ручной сборки» стала заметной и для авторов многих собственно лингвистических исследований. При этом речь идет о следующих проблемах, связанных с объемом, разнообразием и способом разметки текстов в корпусе:

1. собственно недостаточный объем данных;
2. недостаточный объем лингвистически размеченных данных;
3. недостаточный объем и типологическое разнообразие текстов с метатекстовой разметкой по различным релевантным параметрам варьирования (т. н. дифференциальная неполнота [Belikov, Kopylov, 2013]).

Потенциально исчерпывающе полный источник данных — интернет, не является корпусом сам по себе, поскольку не имеет разметки и не дает сколько-нибудь точной статистики даже по поддерживаемым ограниченным видам запросов. Поэтому с середины нулевых годов бурно развивается направление WAC (Web As a Corpus), целью которого является получение автоматически размеченных мега-корпусов из интернета.

Примерами таких корпусов для русского языка являются RuWac [Sharoff, Nivre, 2011] и недавно появившийся, но уже получивший популярность у исследователей, сделанный на основе SketchEngine Адама Килгариффа корпус RuTenTen, содержащий более 10 млрд слов [Jakubíček, 2013] и собственно разрабатываемый авторами ГИКРЯ_1.0 [Belikov, Piperski, 2013].

Все три корпуса используют сейчас одну и ту же процедуру первичной морфоразметки (таггер С. Шарова [Sharoff, Nivre, 2011]). RuTenTen и ГИКРЯ_1.0 являются близкими по объему (хотя в плане развития ГИКРЯ установлен ориентир в 50 млрд слов и более).

Принципиальным отличием ГИКРЯ_1.0, о котором и идет речь в данной статье, является то, что два других интернет-корпуса (RuWac и RuTenTen) получены в результате «слепого» кроллинга русскоязычного интернета, что не позволяет решить проблему дифференциальной полноты. В частности, в них полностью отсутствует метатекстовая разметка. Кроме того, по соображениям эффективности кроллинг в таких корпусах ограничивается ресурсами в домене .ru, представленными текстами в HTML без интерфейса, что исключает многие полезные ресурсы, в том числе и такие важные, как vk.com или blogs.mail.ru.

ГИКРЯ с точки зрения закладываемой в корпус информации отличается тремя важными особенностями:

1. неслучайным способом набора корпуса с учетом сегментной структуры интернета;
2. более жесткой процедурой отбора и очистки страниц, включая их декомпозицию на отдельные объекты анализа в случае структурной неоднородности (например, пост — комментарии);
3. максимально широким использованием априорной метатекстовой разметки, которую можно получить аккуратным анализом источников корпусных данных (прежде всего — разбором профилей авторов).

Эти отличия в построении корпуса оказываются очень существенными в отношении результатов запросов к нему в сравнении с другими автоматическими интернет-корпусами.

Сегменты интернета и система метатекстовой классификации

Было показано [Беликов 2006; 2010 и др.], насколько различаются результаты лингвистического анализа (прежде всего в области лексикографии и лексикализованного синтаксиса), если перейти от усреднения данных к учету распределения авторов текстов по различным параметрам.

При этом в отсутствии адекватных инструментов подобный анализ производился с помощью требовавшего колоссальных усилий «ручного» анализа данных, полученных помощью интернет-поисковиков.

Появление мегакорпусов с метатекстовой разметкой открывает дорогу к получению подобных результатов «легко и непринужденно».

При этом остается открытым вопрос, как добиться дифференциальной полноты данных в интернет-корпусе при наличии большого числа параметров социолингвистического и жанрового варьирования [Belikov V., Kopylov N., 2013].

Если по социолингвистическим и региональным параметрам априорная разметка может быть получена из самих данных, то жанровая полнота требует как минимум ясного понимания устройства жанровых категорий, чего, увы, не наблюдается [Crowston, 2010]. Эта цель может быть достигнута не сразу, а в результате нескольких итераций создания корпуса с параллельно ведущейся работой по созданию надежной системы параметров жанровой разметки и методов автоматической жанровой классификации.

В настоящее время в проекте ГИКРЯ идет такая работа (см. статью [Сорокин, Катинская, 2014] в данном сборнике).

И пока такая система разрабатывается, эффективным способом обеспечить относительную типологическую полноту корпуса на первых этапах его сбора является опора на сегментную структуру интернета. Мы добиваемся полноты сегментной структуры корпуса, коррелирующей с жанровой полнотой.

Сегментом интернета мы называем [Belikov V., Selegey V. 2012] компактное множество страниц Интернета, объединенное некоторой общей коммуникативной целью создателей text-rich контента. При этом диапазон варьирования социолингвистических характеристик авторов внутри таких сегментов может

быть весьма значителен. К таким функционально однородным «авторизованным» сегментам интернета, которые мы используем для сбора первой версии ГИКРЯ относятся:

- Блоги
- Микроблоги
- Форумы
- Социальные сети
- Сайты, аккумулирующие авторизованные тексты художественного и публицистического характера
- Энциклопедические ресурсы в которых имеются авторы статей (а не анонимные группы авторов и редакторов, как в Википедии)
- Новостные ресурсы.

Различие в структуре и принципах отбора сравниваемых интернет-корпусов очевидно из таблиц 1 и 2. Как мы видим, типичный подход создателей интернет-корпуса состоит в достижении некоего усреднения языковой картины с помощью набора данных из максимально большого числа источников (многих тысяч!). В проекте ГИКРЯ подход принципиально отличается: сбор корпуса происходит именно в соответствии с априорной схемой сегментной структуры интернета.

Таблица 1. Доменный состав интернет-корпусов RuWac и RuTenTen

RuWac		**RuTenTen**	
Число документов	Домены	Число документов	Домены
323 300	livejournal.com	114 427	spb.ru
37 860	narod.ru	82 606	narod.ru
8 293	lib.ru	39 649	tomsk.ru
6 117	germany-rest.com.ua	38 383	org.ru
5 966	bibliotekar.ru	34 752	net.ru
5 862	sites.google.com	32 419	gov.ru
5 844	subscribe.ru	22 668	ucoz.ru
5 755	shkolazhizni.ru	19 433	karelia.ru
5 423	lenta.ru	19 227	mos.ru
5 297	russ.ru	18 598	edu.ru
4 814	hotmail.ru	18 372	com.ru
3 602	football.hiblogger.net	17 736	nnov.ru
3 528	yandex.ru	17 623	msu.ru
3 491	org.ua	16 961	perm.ru
3 487	spb.ru	15 597	rospotrebnadzor.ru
3 478	eka-mama.ru	13 986	forum2x2.ru
3 462	mail.ru	13 490	msk.ru
3 171	falppo09.ru	13 228	rfn.ru
3 160	org.ru	12 910	academic.ru
...	...	...	...
Всего: 2 млн		Всего: 35 млн	

Таблица 2. Сегментная структура ГИКРЯ_1.0

Сегмент	Слов (млн.)	Документов
Блоги мейл.ру (комментарии к топикам)	186	~6 млн
Живой журнал	7900	~ 50 млн
В контакте	2000	~100 млн
Журнальный Зал	306	56 тыс.
Новости	700	~ 2 млн
Форум awd.ru	190	~2 млн
Всего	11 282	>100 млн

Комментарии к таблице 2

1. С развитием блогосферы в ней все шире развивается новостной и иной (гороскопы, кулинарные рецепты, «мудрые притчи», анекдоты и т. п.) репостинг, который для лингвиста представляет собой шум. Дублирование записей, нередко достигающее нескольких тысяч, особенно характерно для первичных записей в блогах мейл.ру. Поэтому в версию 1.0 ГИКРЯ с этой платформы включены лишь комментарии. В дальнейшем предполагается неновостной репостинг этой и иных блоговых платформ выделить в отдельные тематические сегменты, что, в частности позволит оценить их поло-возрастную привязку.
2. В первую версию ГИКРЯ не входят в качестве сегментов тексты, представляющие т.н. языки для специальных целей (LSP/ЯСЦ). В узком понимании это языки науки, но при широком термин покрывает и любое профессиональное, и религию с эзотерикой, и самые разные «клубы по интересам», фан-движения и т.п. В той или иной, но сильно разной степени эти языки находят реализацию в интернете. Многое реализуется в «жанре» форума, но имеется и огромное число специализированных сайтов. Обеспечение дифференциальной полноты по «тематическим» параметрам является отдельной и очень сложной задачей, но она не является все же первоочередной с точки зрения исследования языковой вариативности.
3. Также пока не учитываются частично ортогональные различия по социолингвистическим характеристикам типичного адресата. Адресат и адресант в некоторых сегментах идентичны, но есть много сайтов со специализированным адресатом: детским, подростковым, женским (отдельно — *мамским*, как сейчас говорят, то есть адресованных беременным и матерям грудных младенцев).

С точки зрения пользователя корпуса отличие ГИКРЯ состоит в том, что можно задавать сегмент в качестве параметра запроса, реализуя тем самым «как бы жанровое» ограничение на исследуемый материал. В случае RuWac и RuTenTen такое ограничение задать практически невозможно (что в отсутствие и любых других параметров метатекстовой разметки делает задачу получения дифференциальной выдачи невозможной).

Первая версия ГИКРЯ с точки зрения объема дифференцированных данных

Априорная информация об авторах, связанная с отдельными сегментами, дает (в не всегда достижимом идеале) возможность проводить дифференциальный анализ по следующим социолингвистическим параметрам:

- Возраст
- Пол
- Регион
- Образовательный уровень

Объем априорной социолингвистической разметки в ГИКРЯ_1.0 (11 282 млн слов на 1.04.14) представлен в табл. 3.

Комментарии к таблице 3

1. Суммарный объем проиндексированных данных в ГИКРЯ к июню 2014 года составит около 12 млрд слов, что соответствует примерно 20% того объема, который первоначально планировалось иметь в корпусе в конечном итоге. Более точные оценки, связанные с расчетом дифференциальной полноты по релевантным параметрам, можно будет дать несколько позже.
2. Относительно невысокая скорость набора данных в корпус объясняется необходимостью проведения двух операций, требующих участия программистов и лингвистов, анализирующих очередной сегмент интернета:
   - разработки соответствующего метода очистки страниц (удаления обвязки);
   - извлечения данных об авторах
3. Даже при текущем объеме первой версии ГИКРЯ количество документов с априорной разметкой исчисляется десятками миллионов с общим объемом в несколько миллиардов словоупотреблений. На таком объеме материала уже можно основывать серьезные дифференциальные исследования.

Таблица 3

Сегмент	Пол	Возраст	Регион
Блоги мейл.ру	164	81	113
Живой журнал	0	1800	5600
Вконтакте	2000	1600	1600
Новости	0	0	0
Журнальный зал	258	0	0
Форумы adw.ru	163	0	0
Всего:	2585	3481	7313

О надежности априорной разметки

Априорная разметка в некоторых сегментах корпуса безусловно не является абсолютно надежной. Она колеблется для разных сегментов интернета в диапазоне от 85 до 95 % в зависимости от исследуемого параметра.

Исследования по автоматической социолингвистической разметке различных социальных сетей, и отдельные работы, посвященные оценке достоверности авторских данных (прежде всего, в Twitter, например [D. Nguyen, 2013]) показывают, что имеются существенные систематические факторы, сдвигающие такие данные. Кроме того, не является очевидным, на каких шкалах нужно производить такие оценки в случае выставления автоматических и смешанных признаков в корпусе. В частности, имеются аргументы в пользу небинарных гендерных шкал [J. Lorber, 1996].

В этой статье мы не будем касаться вопросов, связанных с анализом процессов «самопозиционирования» авторов интернета, а также сравнением результатов априорной и автоматической социолингвистической и региональной атрибуции авторов текстов.

Для целей этой статьи достаточно указать, что имеется достаточно высокая корреляция между реальными данными (насколько они извлекаются из самих текстов), авторской самоидентификацией и результатами автоматической классификации.

В целом это позволяет изучать как собственно объективные корреляции между языковыми и социолингвистическими параметрами, так и систематические девиации между прогнозируемыми и априорными характеристиками.

Таким образом, использование априорной разметки дает достаточно надежные результаты (с погрешностью в единицы процентов). При возрастном анализе языка блогосферы для единиц, употребление которых существенно зависит от возраста, у лиц 12–69 лет «статистические результаты обычно хорошо укладываются в „правильную“ картину, что позволяет предполагать серьезное преобладание здесь тех, кто указал фактический возраст. ‹…› При анализе возрастного распределения сниженной лексики возраст, начиная с которого получаемые данные становятся недостоверными, снижается» [Беликов 2012], в таких случаях использовать данные о лицах старше 60 нецелесообразно.

Примеры дифференциального анализа запросов в ГИКРЯ по различным параметрам

Гендерное варьирование

В последнее время гендерному анализу социальных сетей уделяется большое внимание. Можно выделить три основных направления:

- гендерная лингвистика (просто гендерная вариативность)
- гендерная психолингвистика (попытки интерпретации)
- гендерная реклама: тут важны не собственно лингвистические отличия, но любые признаки, позволяющие выявить индивидуальные предпочтения пользователя.

Последнее направление преобладает, поскольку его продвигают рекламодатели. Гендерная атрибуция основывается на побочных признаках, но зато пополняет массив атрибутированных текстов.

Журнальный сегмент ГИКРЯ (ЖС) содержит все собственно журнальные тексты Журнального зала (ЖЗ) по состоянию на апрель 2014 г. (56 тыс. текстов, 306,0 млн словоупотреблений); в этом сегменте допускается классификация по конкретным изданиям, годам публикации и создателям (создателем текста считается его автор или переводчик).

В дальнейшем предполагается пополнять ЖС новыми публикациями в ЖЗ, текстами тех изданий, которые на собственных сайтах представлены полнее, чем в ЖЗ, а также систематически включать доступные в интернете в оцифрованном виде публикации толстых журналов, не входящих в ЖЗ, как «центральных» («Москва», «Юность» и др.), так и региональных — «Вологодская литература» (Вологда), «Дарьял» (Владикавказ), «Бельские просторы» (Уфа), «Дальний Восток» (Хабаровск) и др.

В настоящее время главной задачей обработки этого сегмента является подготовка полноценных профилей создателей текстов (пол, год рождения, региональная привязка), что связано со значительным объемом ручной работы. В ходе тестовых поисков по ЖС наиболее интересными оказались результаты анализа гендерных различий в узусе.

Поиск с учетом пола проводится на 84,3 % текстов ЖС, общий объем мужских словоупотреблений — 194,2 млн, женских — 63,8 млн. Мужской подкорпус превышает женский в 3,04 раза, то есть соотношение гендерно нейтральных единиц должно быть близко к 3.

На этом подкорпусе тестировалась гендерная предпочтительность различных дискурсивных слов, коллокаций, фразеологизмов. Невысокие абсолютные цифры не позволяют делать серьезных выводов, но там, где выдача составляет тысячи вхождений, различия явно достоверны, ср. «феминизированность» выражения *каждый раз* и «маскулинизированность» по *меньшей мере* в Табл. 4:

Таблица 4

Выражение	**M**	**F**	**M/F**
стерпится слюбится	31	13	2,4
как снег на голову	105	44	2,4
каждый раз	**3657**	**1475**	**2,5**
как заведенн(ый/ая/ые)	224	84	2,7
всего лишь	5954	1998	3,0
не на шутку	648	219	3,0
наверняка	4931	1667	3,0
по крайней мере	6819	2178	3,1
как правило	5291	1711	3,1
авось (кроме *на авось*)	670	209	3,2
в несколько раз	419	131	3,2

Выражение	**M**	**F**	**M/F**
небось	1622	498	3,3
в полном разгаре	68	20	3,4
по меньшей мере	**1533**	**438**	**3,5**
почем зря	225	54	4,2

Остановимся детальнее на двух случаях гендерного противопоставления в узусе.

1. Выявились различия при описании количества: импрессионистическое описание (типа *очень много*, *так много*) почти нейтрально, хотя несколько более свойственно женщинам, а сопоставительное (типа *заметно больше*, *в … раз больше*) явно оказывается более мужским, ср. данные в Табл. 5:

Таблица 5

	муж.	**жен.**	**муж./ жен.**
очень много	3278	1090	3,01
так много	3204	1267	2,53
раз больше	536	110	4,87
раза больше	460	107	4,30
чересчур много	115	23	5,00
существенно больше	63	20	3,15
заметно больше	38	10	3,80

Абсолютные цифры в последних трех строках невелики, и их пока не стоит принимать во внимание, но гендерные различия между *очень/так …* и *в … раз* вполне очевидны и подтверждаются аналогичными конструкциями с другими наречиями, ср. отношение мужских словоупотреблений к женским в Табл. 6:

Таблица 6

		много	**мало**	**Высоко**	**низко**	**быстро**
1	*очень*	3,0	2,8	2,8	2,7	2,5
2	*так*	2,5	2,3	2,8	2,3	2,6
5	*раз(а)* + **comp**	4,6	4,7	3,3	3,3	4,4

2. Имя *Кондратий* нередко ассоциируется со смертью; нет сомнений что восходит этот факт к фразеологизму, этиология которого иногда возводится к Кондратию Булавину. Отвлекаясь от деталей, можно констатировать, что исконно во фразеологизме имя употреблялось в уничижительном варианте *кондрашка*, а семантика была привязана к параличу. Ученые выявили точное значение фразеологизма (‘скоропостижно умереть, скончаться’) и его форму — *кондрашка* сочетается с глаголом *хватить* в прошедшем времени

(порядок компонентов и род глагола не фиксированы), а Партия и Правительство утвердили такую норму при использовании русского языка в качестве государственного[1].

Между тем литераторы (как и иные носители русского языка) продолжают использовать во фразеологизме полную форму имени, а также инфинитив глагола; имя может использоваться и вне фразеологизма как персонификация явления, ср.: *Кондратий, как говорят, у всех за левым плечом* (Ирина Богатырева, «Приступ»); *От всех пережитых волнений в новогоднюю ночь у меня случился инсульт. Кондрашка. Удар* (Эдуард Русаков, «Валерик»).

Статистика употребления имени *Кондратий/Кондрашка* в таком значении отдельно и в сочетании с глаголом (*с*)*хватить* в Журнальном зале приведена в Табл. 7:.

Таблица 7

пол автора	**Имя *кондр…* как нарицательное (всего)**		**Фразеологично со (*с*)*хватить***	
	кондрашка	*кондратий*	*кондрашка*	*кондратий*
муж.	**62**	43	**54**	30
жен.	9	**15**	7	**11**

Несмотря на невысокие цифры, предпочтение мужчинами уничижительного варианта, а женщинами — полного выглядит достаточно убедительно.

Региональное варьирование

В предисловии к словарю «Языки городов» [2008] говорилось: «в Приуралье и Сибири *уколы* и *прививки* часто не *делают*, а *ставят*, то есть у слов *укол* и *прививка* есть региональная специфика»; этот вывод делался на основании газетных материалов базы СМИ «Интегрум», где на август 2007 г. в Урало-Сибирском регионе имелось 828 текста с сочетанием (*по*)*ставить укол* при 772 соотношение текстах с глаголом (*с*)*делать*. На периферии ареала — в Казахстане и но Дальнем Востоке глагол ставить употреблялся в этом контексте в пять раз реже, чем делать, а у ближайших западных соседей разница увеличивалась до 52 раз.

ГИКРЯ позволяет соотнести данные газетных текстов с повседневным словоупотреблением и узусом профессиональных литераторов.

Повседневный узус тестировался в Живом журнале по записям с этими глаголами и уколом в одном предложении[2]. В Урало-Сибирском регионе в целом соотношение глаголов (*по*)*ставить* и (*с*)*делать* в этом контексте составляет

---

[1] Подробнее см. Беликов 2010-b.

[2] Шум типа *делает уколы, ставит банки* на Урале и в Сибири в обоих типах выдачи распределяется относительно равномерно, в прочих регионах он завышает статистику на (*по*)*ставить укол*.

1:0,9[3]; лишь в характеризующихся трудовой иммиграцией автономных округах Тюменской области наблюдается двукратное превосходство глагола (*с*)*делать*. Таково же соотношение этих контекстов на Дальнем Востоке и в Казахстане. В ближайших западных субъектах федерации суммарное соотношение контекстов различается в шесть раз.

В ЖС проводился только поиск контекстов с глаголом (*по*)*ставить*[4]. *После отсеивания шума, который включает и контексты, где укол* и (*по*)*ставить* синтаксически связаны (ср. *ставить ей в счет все булавочные уколы*; *поставили на ноги каким-то уколом*), нашелся 81 релевантный текст 68 авторов. Среди них лишь у 13 не устанавливается явная связь с урало-сибирским ареалом.

У литераторов связанных с ареалом рассматриваемой конструкции, такое употребление глагола (*по*)*ставить* явно не имеет стилистических ограничений, встречается в дневниковых записях, драматургических ремарках и т. п. Для наиболее публикуемых авторов проводилось сопоставление частотности в этом контексте обоих глаголов. У Н. Горлановой (Пермь) в 12 публикациях (часть — в соавторстве с В. Букуром) каждый из них встретился по 8 раз, у Э. Русакова (Красноярск) в шести публикациях — трижды (*с*)*делать*, пять раз (*по*)*ставить*. Характерная для региональной нормы синонимия глаголов позволяет избегать нелюбимых отечественной стилистикой тавтологических повторов, ср.: ‹…› *если бы не Шура, которая насильно делает ей уколы, она бы давно уже спокойно умерла. А Шура ставит и ставит уколы* — Р. Солнцев (Красноярск), «Старица»; *Никто так и не понял, что десятиклассники делали со шприцами, — уж точно не уколы себе ставили, это было бы абсурдно, ведь никто не любит уколы* — Андрей Юрич (Якутия/Кемерово), «Ржа».

Выводы и обсуждения

Нужны или не нужны дифференциальные корпуса — ответ на этот вопрос не является самоочевидным. Возможно, для каких-то задач полезно тотальное усреднение данных на максимально больших массивах языковых данных, что хорошо укладывается в идею RuTenTen. Известны и другие стратегии «нормализации», например, проведение языкового анализа на материале статей Википедии, в которых в теории многослойное редактирование вымывает индивидуальные характеристики (в русской Википедии, впрочем, немало опечаток и явных грамматических ошибок, много неудачных переводов и калек с английского, но встречаются даже с украинского).

Но результаты наших исследований показывают, что при изучении языковых конструкций дифференциальные особенности в употреблении могут обнаруживаться не только там, где они интуитивно ожидаются.

---

[3] Нет данных по Туве.

[4] Выяснение того, как часто литераторы Урала и Сибири пользуются в этом контексте глаголом (*с*)*делать*, возможно лишь после создания полноценных профилей для всех авторов.

Проект ГИКРЯ переходит из стадии «корпусного строительства» в стадию экспериментального использования: разработчики не готовы пока открыть корпус для свободного пользования всем желающим, но все заинтересованные исследователи могут получить доступ к нему на условиях участия в тестировании.

Литература

1. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.* Corpus as language: from scalability to register variation In Proc. Int. Conf. on Computational Linguistics "Dialog", 2013
2. *Belikov V., Piperski A., Selegey V., Sharoff S.* Big and diverse is beautiful: A large corpus of Russian to study linguistic variation (in co-authorship with V. Belikov, A. Piperski, S. Sharoff) — In Proc. of the 8th Web as Corpus Workshop (WAC-8) / Corpus Linguistics Conference 2013
3. *Belikov V., Selegey V., Sharoff S.* Preliminary considerations towards developing the General Internet Corpus of Russian. — In Proc. Int. Conf. on Computational Linguistics "Dialog", 2012
4. *Crowston, K., Kwasnik, B., Rubleske, J.* 2010. Problems in the use-centered development of a taxonomy of web genres. // Mehler, A., Sharoff, S., Santini, M., editors, Genres on the Web: Computational Models and Empirical Studies. Springer, Berlin/New York.
5. *Jakubíček Miloš, Kilgarriff Adam, Kovář Vojtěch, Rychlý Pavel, Suchomel Vít.* The TenTen Corpus Family // Int Conf on Corpus Linguistics, Lancaster, July, 2013.
6. *J. Lorber.* 1996. Beyond the binaries: Depolarizing the categories of sex, sexuality, and gender. Sociological Inquiry, 66(2): 143–160.
7. *D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder.* 2013. "How old do you think I am?" A study of language and age in Twitter // Proceedings of ICWSM 2013
8. *Rosenthal S. and McKeown K.* 2011. Age prediction in blogs: a study of style, content, and online behavior in pre- and post-social media generations // Proceedings of ACL 2011.
9. *Sharoff S., and Nivre, J.* The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. // Proc. Dialogue, Russian International Conference on Computational Linguistics, Bekasovo, 2011.
10. *Беликов В. И.* 2006. Словарь «Языки русских городов»: подбор примеров и интернет // «Компьютерная лингвистика и интеллектуальные технологии». Труды Международной конференции Диалог 2006. М.: Ин-т проблем информатики РАН, 2006.
11. *Беликов В. И.* 2010-а. Методические новости в социальной лексикографии XXI века // Slavica Helsingiensia 40. Instrumentarium of Linguistics. Sociolinguistic Approaches to Non-Standard Russian, Helsinki, 2010 A. Mustajoki, E. Protassova, N. Vakhtin (eds.).

12. *Беликов В. И.* 2010-b. О словарях, «содержащих нормы современного русского литературного языка при его использовании в качестве государственного языка Российской Федерации» // Грамота.Ру [http://gramota.ru/biblio/research/slovari-norm/].
13. *Беликов В. И.* 2012. К методике корпусного исследования лексики. Рукопись. (Для сборника Русский язык и новые технологии. Сост. Г. Ч. Гусейнов. М.: НЛО.)
14. Языки городов. 2008. Материалы к словарю региональной лексики. В составе электронного издания: ABBYY Lingvo X3 ME: CD. M.: ABBYY, [http://www.lingvo.ru/goroda/].

References

1. *Sharoff, S.,* 2007. Classifying Web corpora into domain and genre using automatic feature identification. // Proc. of Web as Corpus Workshop, Louvain-la-Neuve.
2. *Sharoff S.,* 2007. Creating General-Purpose Corpora Using Automated Search Engine Queries.
3. *Беликов В. И., Селегей В. П., Шаров С. А.*, 2012. Пролегомены к проекту Генерального интернет-корпуса русского языка. // Труды конференции Диалог 2012.
4. *D. Bamman, J. Eisenstein, and T. Schnoebelen*. 2012. Gender in Twitter: styles, stances, and social networks. CoRR.
5. *M. Ciot, M. Sonderegger, and D. Ruths. 2013.* Gender inference of Twitter users in non-English contexts. In Proceedings of EMNLP 2013.
6. *C. Fink, J. Kopecky, and M. Morawski.* 2012. Inferring gender from the content of tweets: A region specific example. In Proceedings of ICWSM 2012.
7. *S. Goswami, S. Sarkar, and M. Rustagi.* 2009. Stylometric analysis of bloggers' age and gender. In Proceedings of ICWSM 2009.
8. *A. Mukherjee and B. Liu.* 2010. Improving gender classification of blog authors. In Proceedings of EMNLP 2010.
9. *C. Peersman, W. Daelemans, and L. Van Vaerenbergh.* 2011. Predicting age and gender in online social networks. In Proceedings of SMUC '11.
10. *D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta.* 2010. Classifying latent user attributes in Twitter. In Proceedings of SMUC 2010.
11. *S. Rosenthal and K. McKeown.* 2011. Age prediction in blogs: a study of style, content, and online behavior in pre- and post-social media generations. In Proceedings of ACL 2011.

USING DISTRIBUTED REPRESENTATIONS FOR ASPECT-BASED SENTIMENT ANALYSIS

Blinov P. D. (blinoff.pavel@gmail.com),
Kotelnikov E. V. (kotelnikov.ev@gmail.com)

Vyatka State Humanities University, Kirov, Russia

The article is focused on aspect-based sentiment analysis, which is a specific version of the general sentiment analysis task. Its goal is to detect the opinions expressed in the text on the level of significant aspects of the specified entity. An overview of the existing approaches and previous work is presented.

The main result of our work is a new method of aspect-based sentiment analysis based on the distributed representations of words. Such representations are obtained by using deep learning algorithms. The method includes the well-known algorithm of training distributed representations of words, two new techniques for constructing the aspect and sentiment lexicons, and an algorithm for calculating aspect scores.

Examples of aspect and sentiment terms are given. The vectors of resulting terms are visualized using the t-SNE method. The article presents the results of experiments on a test corpus for three aspects—"food", "interior" and "service", which yield aF1-measure increase of 11 to 16% as compared to the baseline.

Key words: aspect-based sentiment analysis, machine learning, deep learning, distributed representations

1. Introduction

The area of sentiment analysis is actively developing recently. The sentiment analysis is the problem of finding user opinions and sentiments in a text [10]. The problem is evidently still far from its final solution therefore it is interesting in the academia. The new methods of computational linguistics and machine learning are being developed to solve the problem of sentiment analysis. The business community is interested in the commercial applications of such analysis, for example, the sentiment analysis may be useful in the study of opinions and preferences of the target audience of consumers.

The aspect-based sentiment analysis is a relatively new task in this area [10, p. 58]. Its appearance can be explained by the fact that the sentiment analysis on the level of a whole text or even on the sentence level is not able to detect the expressed opinion on the certain aspects of the studied entity. Such formulation of the problem saves the common sense of the sentiment analysis and at the same time it is more detailed and researches the opinions expressed in the text on the level of the meaningful aspects

of an entity. For example, the sentence “In general, the food is great, but the service is terrible!” presents different opinions on aspects “food” and “service” of a single object “restaurant”. Along with the sentiment terms detection task such version of sentiment analysis is extended with the aspect extraction task [10, p. 67].

An aspect term can be defined as a word or a collocation that explicitly determines an attribute of a target entity. A sentiment term is a word or a collocation that expresses the user’s subjective opinion. Both types of terms, aspect and sentiment, vary from one domain to another, therefore the development of the effective methods of automatic selection of aspect and sentiment terms with minimal time costs and human labor is very important.

A continuous vector space of distributed representations [17] of words as the source of lexicons constructing is investigated in the article. The new techniques for the aspect and sentiment lexicons constructing from small initial sets of words are proposed.

The remainder of the article is as follows. The overview of the previous approaches and papers is given in section 2. Section 3 describes the used corpus of documents. The techniques for aspect and sentiment terms detection are presented in section 4. The results of the experiments and the conclusions are given in sections 5 and 6 respectively.

2. Related Works

The main two subtasks which must be solved to perform the aspect-based sentiment analysis are the aspect extraction and the sentiment terms detection.

The aspect extraction task can be solved within three main approaches [10, pp. 67–78]:

1. the frequent-based approach;
2. the supervised machine learning approach;
3. the unsupervised machine learning approach.

The core idea of the first approach is to select the most frequent nouns and collocations as the aspect terms [5, 15]. Despite its relative simplicity the approach can show not a bad quality of aspect extraction, however it has some shortcomings: it gives too many false aspect terms and tends to miss infrequent terms. Besides that, the clustering by aspect categories has to be done for the obtained terms.

The aspect extraction task can be expressed in the terms of information extraction task which, as it is known, can be solved by the supervised machine learning methods [6, 7]. The main shortcoming of such approach is a high complexity of obtaining the labeled train data. The result of this shortcoming is the problem of resetting of the methods for the new domains.

The method proposed in the article belongs to the third approach—unsupervised machine learning, which overcomes the mentioned disadvantages of the two previous approaches. The main methods of this approach are the methods of topic modeling for example Latent Dirichlet Allocation (LDA) [1].

As the base LDA model can find only global topics of a document's collection, various modifications of this model which can find the distinct aspects were proposed [9, 19]. The results of the work of such models are the probability distributions on the words, which correspond to the aspects, that is the separation between aspect and sentiment terms is not performed. In our work such separation is performed explicitly what gives a user more interpretability over the result of the analysis.

The LDA method was also used in [2]: the aspect terms were found first and then the sentiment terms which can only be the adjectives were detected. Our method preserves the sequence of the actions, but as the sentiment terms beside adjectives it also takes into account the complex phrases, which are good indicators of sentiment and make analysis more precise.

The paper [12] investigated the method of aspect term generating based on the semi-supervised modeling. For each aspect the initial set of words was specified and then it was replenished with the new terms by using the LDA model. In our work the source of the new terms is the space of continuous vector representations of words, which is obtained by using the deep learning. This approach is more flexible: comparing with the LDA it gives the intermediate representation for each term. The vector space brings the notion of similarities between words, which is useful for solving natural language processing tasks [4, 18, 21].

Another important subtask—the sentiment terms detection—is often solved with the help of sentiment lexicons. Such lexicons list emotionally-colored lexical units and their weights. The main obstacle in using lexicons is the complexity of their creation. They are constructed either manually by the experts or automatically from the initial set of words with their sentiment weights [20, 23]. In [23] the authors used only one initial word ("*good*") and 6 negations for the lexicon creation. In [20] the initial sets consisted only of two words ("*excellent*" and "*poor*"), the sentiments of another phrases were calculated on the base of mutual information measure. Similarly our method of lexicon creation uses two initial words "*отличный*" (*great*) and "*ужасный*" (*terrible*) however the cosine similarity is used to detect the sentiment terms and to calculate their weights. In [20] the author used the search queries as the source of statistical information about terms co-occurrence. In our work such source is the large corpus of documents.

3. The Text Corpus

Unfortunately, there is no available text corpus in Russian because the task of aspect-based sentiment analysis is relatively new so the new corpus was created, it includes the user reviews of restaurants. 33,243 reviews were collected from *restoclub.ru*. For each review the user specifies the numeric score for the following aspects: *food*, *interior* and *service*. Initially the scores were presented in ten-point scale, we cast the scores to the binary scale by the following mapping scheme: {*1..5*}→*negative*, {*6..10*}→*positive*. 15,285 reviews[1] were selected as a test data set, for each review in this set at least one

[1] Test corpus and dictionaries are available at: http://goo.gl/NhEvWu.

of the aspect scores is less than 8. Such selection is made to reduce the imbalance of the collection to the positive scores. The distribution of the test data set in aspects and scores is shown in Table 1 (a single review can have positive score for one aspect and negative score for another).

Table 1. Some statistics of the test data set

Aspect	Positive reviews	Negative reviews	Total
Food	10,063	5,222	15,285
Interior	11,296	3,989	15,285
Service	8,707	6,578	15,285

To build the high quality distributed representations of words we need only the texts of reviews. The quality of the received vectors depends on the quantity of texts, so 14,058 reviews without any aspect scores were additionally collected from *restoran.ru*. Thus, the corpus of 47,301 reviews in total was used to build the distributed representations of words. Note that the aspect scores were not used in this process.

The text of each review was preprocessed with the segmentation, the tokenization and the morphological analysis. The procedures were performed using such tools as Mystem [13] and FreeLing [14].

4. The Aspect-Based Sentiment Analysis

In our work the aspect-based sentiment analysis includes four stages. On the first stage the vector space of distributed representations is built. On the second and the third stages the aspect and sentiment terms are determined respectively. The scores for each aspect are calculated on the final stage.

4.1. The Vector Space

The unsupervised deep learning algorithms were used to build the vector space. The common idea of such algorithms is to automatically find the "good" set of features to represent in high quality the target object (image, audio signal, text, etc.).

In case of textual information each lexical unit (word) is represented by the vector of real numbers called *distributed representation* [17]. The peculiarity of such representations is that they encode a set of degrees of linguistic similarity between words. In other words, semantically and syntactically related words appear together in the vector space.

To build such space the skip-gram model was used [11]. Formally, the model tries to maximize the following function for the given train sequence of words $w_1, \ldots, w_T$ [11]:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p(w_{t+j} \mid w_t) \to \max, \tag{1}$$

where c—the size of the training context (window size), T—the length of the train sequence of words.

The probability $p(w_{t+j} \mid w_t)$ is defined as [11]:

$$p(w_O \mid w_I) = \frac{\exp({v'_{w_O}}^T v_{w_I})}{\sum_{w=1}^{W} \exp({v'_w}^T v_w)}, \quad (2)$$

where v_w and v'_w are the input and output vector representations of w; w_I and w_O are the current and predicted words, W—the number of words in vocabulary.

For the experiments we used the Gensim [16] implementation of the skip-gram model. All texts of the corpus (47,301 reviews) presented as a single sequence of sentences were used to build the vector representations of words. On the base of this corpus we construct the lexicon with the words which frequency is more than 5. Next, the dimensionality of the space is chosen (in our case 150). The greater number of dimensions allows to capture more language regularities but leads to more computational complexity of the learning. Each word from the lexicon is associated with the real numbers vector of the selected dimensionality. Originally all the vectors are initialized with random numbers close to zero. During the learning procedure the algorithm "slides" with the fixed size window (in our case 5) along the words of the sequence and calculates the probability (2) of context words appearance within the window on the base of its central word under review (or more precisely, its vector representation). The ultimate goal of the described process is to get such vectors for each word, which allow to predict its probable context. This goal is achieved by maximizing the function (1).

4.2. The Aspect Lexicon Construction

The idea of our method is to extend automatically the initially specified sets of terms. Five initial terms were selected for each aspect (Table 2). The assumption was made that the aspect terms can only be the single words.

For each term in the vector space of distributed representations we can find its nearest neighbors. The cosine similarity was used as a measure of similarity between vectors. Formally, the similarity between two vectors $\vec{a} = (a_1,...,a_n)$ and $\vec{b} = (b_1,...,b_n)$ is given by:

$$similarity(\vec{a},\vec{b}) = \cos(\theta) = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \cdot \sqrt{\sum_{i=1}^{n} b_i^2}}, \quad (3)$$

where Θ—the angle between the vectors, n—the dimensionality of the space.

Thus a list of several closest terms can be formed for each term. In our experiments we took 10 of such nearest terms. By joining all the lists and excluding the repeated terms a set of new terms emerges. We call such new set of terms a *generation*. The initial set of words can be considered as the zero generation. The repeating of the same procedure for the new generations is the iterative process which generates the

aspect terms. The noise words can appear in a generation, especially when the number of iteration is getting large. To keep the thematic coherence of terms under control the additional constraint was injected: at least three of five terms from the zero generation must be close (*similarity* > 0) to the new term. After some of such iterations the set of possible terms runs out and the whole process terminates. Finally, the aspect terms vocabulary is formed by joining all the generations. In our experiments after 10 iterations the lexicon of 3,080 terms was formed. It included 1,749 terms for aspect "*food*", 996 terms for "*interior*" and 335 for "*service*". Table 2 shows the terms of the zero and the first generations. Note that the terms are given in its original spelling. The capability to find such low-frequency terms appears due to the specifics of the vector space of distributed representations.

Table 2. The aspect terms of the zero and the first generations

Food	Interior	Service
Generation 0 (initial sets)		
закуска, суп, десерт, салат, плов	интерьер, атмосфера, музыка, дизайн, бар	обслуживание, персонал, официант, менеджер, сервис
Generation 1		
оливье, солянка, шашлык, похлебка, штрудель, салатик, сметанник, манта, фрикаделька, блюдо, соление, закусочка, явство, ассорти, хачипури, люля, щи, морепродукт, хинкали, хачапури, чизкейк, баранина, нарезка, цезарь, тортик, мороженое, медовик, эклер, уха, супчик, кебаб, сациви, …	саксофон, убранство, стилистика, комфортность, музыкант, беззаботность, гитара, интерьр, уют, обстановка, lounge, продуманность, вокал, репертуар, отмосфера, клуб, интереьер, dj, комфорт, джаз, исполнитель, диджей, времяпровождение, звук, кабак, оформление, …	заместитель, обслужевание, бармен, управлять, девушка, коллектив, сотрудник, официант, официатка, администратор, тамада, внимательность, официантка, директор, обсуживание, команда, девочка, обслуга, отзывчивость, еда, официантка, дирекция, …

Using the t-SNE (t-Distributed Stochastic Neighbor Embedding) [22] algorithm the results of the process can be visualized on the plain. Figure 1 shows the aspect terms vectors for the first four generations. One can trace the cluster structure according to the aspects.

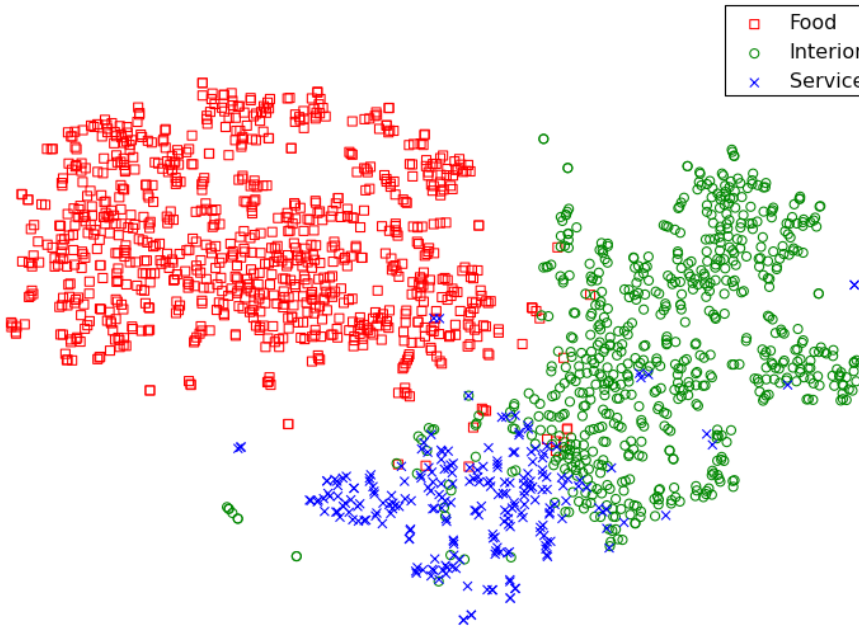


Fig. 1. The aspect terms for the first four generations

4.3. The Sentiment Lexicon Construction

The method of sentiment lexicon construction includes two steps: candidates' selection and its weighting.

As the sentiment phrases can consist of more than one word the additional preprocessing is required. It is known that the modifiers and the negations have the significant meaning in sentiment phrases. The possible set of such words for Russian is proposed in [8]. By our estimates the adverb "*очень*" (*very*) bears the most amplifications and the particle "*не*" (*not*) covers the most negations. By simple pattern

$$<very \mid not> + <very \mid not> + <adjective \mid verb \mid adverb>$$

complex lexical units were formed, for example, *не_готовый, очень_сытный, очень_не_приятный, не_очень_опрятный,* etc. Of course such a way doesn't take into account the whole variety of sentiment phrases, but definitely covers the essential part of it.

We took the single adjectives and the set of complex lexical units as the candidates to the sentiment lexicon. It was the list of $N = 7312$ such candidates, about 34% of them were complex lexical units.

Besides the thematic similarity in the vector space of distributed representations the emotional similarity between the terms can also be traced. So the space can be used not only for sentiment terms detection, but also as a source of sentiment terms weighting.

For the initial setting of sentiment values the etalon terms were determined: *great* was used for the positive sentiment and *terrible* was used for the negative sentiment. For each candidate from the list two values of similarity (3) with the etalon terms were considered as its weights.

Some examples of the most positive and negative terms obtained in such a way are shown in Table 3 (in their original spelling).

Table 3. The examples of the sentiment terms

Positive	Negative
хорошоя, замечательный, великолепный, превосходный, очень_гостеприимный, прекрасный, велеколепный, дружественный, очень_веселый, очень_хороший, шикарный, доброжилательный, очень_душевный, чудесный, суперский, хороший, приятный, не_пошлый, професиональный, очень_теплый, классный, очень_доброжелательный, супер, очень_дружелюбный, тактичный, безупречный, …	отвратительный, безобразный, очень_плохой, отвратный, ужастный, плохой, хамский, невнимательный, не_заслуживать, очень_обидно, отстойный, не_довольный, ужасный, не_вкусный, нулевой, откровенный, не_очень_позитивный, бездушный, не_ровный, не_способный, безответственный, недопустимо, очень_не_понравиться, дурной, очень_разочаровывать, пренеприятный, ужасно, гадкий, не_выдаваться, …

Similarly to the aspect terms the vectors of sentiment phrases can be plot with the t-SNE method. Figure 2 shows the subset of the most positive and the most negative phrases.

4.4. The Aspect Score Calculation

On the final stage it is necessary to get sentiment scores for each aspect. Every sentence is segmented by the following set of punctuation marks: {*? ! , . : ;*}. For each segment the aspect and sentiment terms from the lexicons are found. For every aspect term the summarized similarity (3) with the zero generation terms is calculated and the maximum value $similarity_{\max}$ is chosen. Then the summarized score sum_{em} of the sentiment terms from the current, the couple of the previous and the couple of the next segments is calculated. The final score of a sentence S is found for each aspect a as follows:

$$S_a = \sum_{a \in A} similarity^{a}_{\max} \cdot sum_{em}, \tag{4}$$

where A—the set of the aspect terms for the sentence.

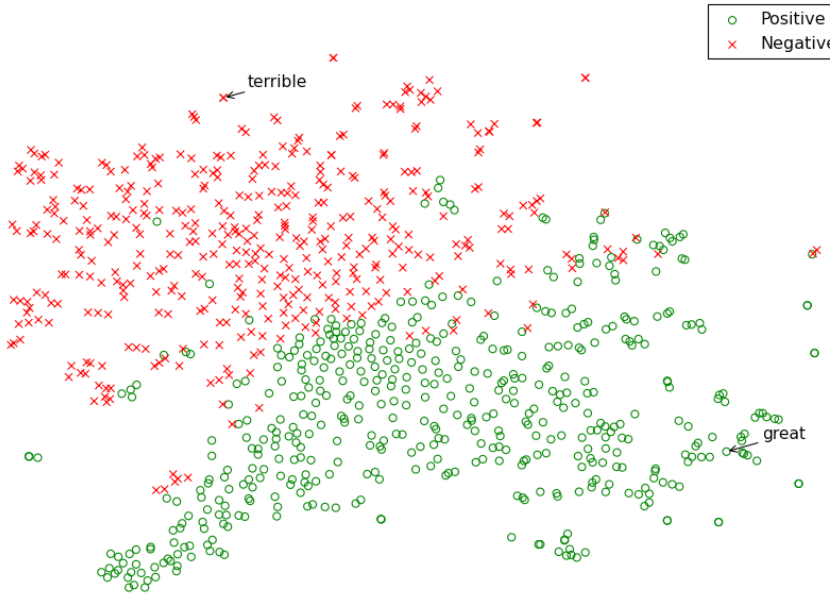


Fig. 2. Some sentiment phrases

The review's emotional score for each aspect a is the sum of scores S_a for every sentence. A sign of this score defines the aspect sentiment—positive or negative.

Table 4 lists some sentences and shows the sentiment calculation for them. The aspect terms are in bold, the sentiment terms are in italics. The numeric values represent either the summarized similarity $similarity_{\max}$ with the zero generation aspect terms or the sentiment score of the term.

Table 4. The examples of the aspect sentiment detection

Предложение	**Оценка**
+1.000 1.482 *Отличный* имбирный **лимонад** с достаточным количеством льда и палочками тимьяна.	$1.482 \cdot 1.000 = +1.482 \Rightarrow pos$
−0.195 1.199 Его мариновали с *какими-то* **травами**, −0,274 которые абсолютно *не понравились*.	$1.199 \cdot (-0.195 - 0.274) = -0.562 \Rightarrow neg$
+0.790 2.046 +0.435 +0.591 А теперь о *прекрасном* =) **Бургер** был *просто чудесным*!	$2.046 \cdot (0.790 + 0.435 + 0.591) = +3.716 \Rightarrow pos$
+0.229 1.737 +0.142 В *общем*, **оформление** *симпатичное*, но я люблю другое.	$1.737 \cdot (0.229 + 0.142) = +0.644 \Rightarrow pos$
−0.431 2.064 Пришёл совершенно *неквалифицированный* **сотрудник**!	$2.064 \cdot (-0.431) = -0.890 \Rightarrow neg$

5. Experimental Results

The proposed method was evaluated according to the precision, the recall and the F1-measure [3]. Figure 3 shows the dependence of the values of the F1-measure for each aspect and the number of the aspect terms (the logarithmic scale). The start values in the baseline correspond to the case when all the scores for the aspect are randomly assigned.

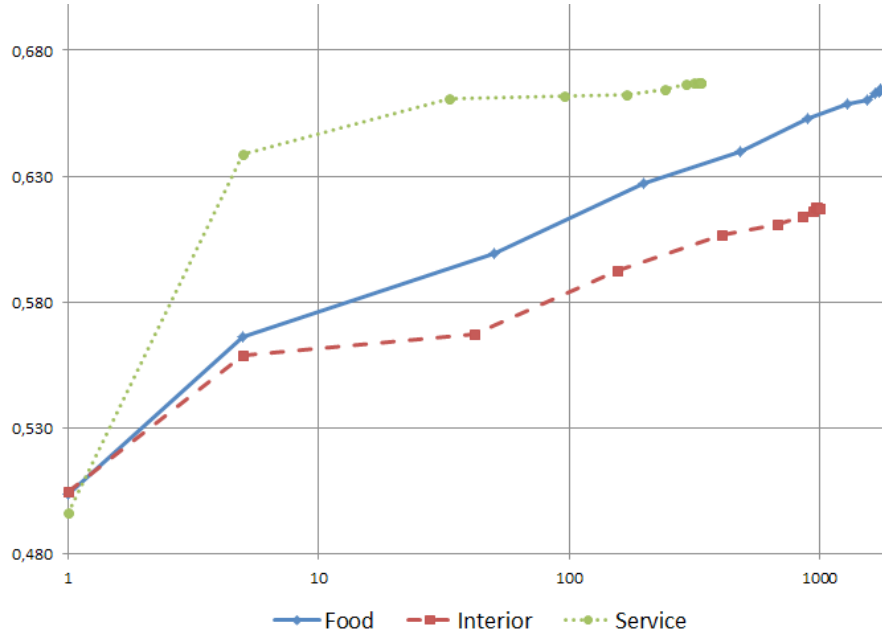


Fig. 3. The F1-measure for each generation

The baseline metrics and the best results of our method (in bold) are shown in Table 5.

Table 5. The evaluation results for the aspects

Aspect	Number of terms	Precision		Recall		F1-measure	
Food	1,749	0.503	**0.686**	0.504	**0.644**	0.503	**0.664**
Interior	996	0.504	**0.629**	0.505	**0.606**	0.504	**0.617**
Service	335	0.497	**0.692**	0.496	**0.644**	0.496	**0.667**

The greatest number of terms is in the "*food*" aspect, because there is a large variety of dish names. When the vocabulary of such terms grows, the value of the F1-measure only increases. In contrast, the aspect "*service*" contains not so many terms and quite a small vocabulary is already sufficient to get almost maximum values of metrics

that were achieved. Low metrics for the aspect "*interior*" can probably be explained by the significant imbalance of the test collection to the positive scores.

6. Conclusion

The article studies the aspect-based sentiment analysis task. The new method of the aspect-based sentiment analysis based on the continuous vector space of the distributed representations is proposed. The suggested method allows to conduct the sentiment analysis with the use of minimal additional information and with minimal dependency from a domain.

The corpus of users' reviews of restaurants is prepared for the experiments. The method is evaluated on this corpus for tree aspects. The result values of the F1-measure significantly outperform the chosen baseline: 66% versus 50% for the aspects "*food*" and "*interior*", 62% versus 50% for the aspect "*service*".

Sentiment lexicon decomposition by aspects seems to be a promising direction to boost the results of the aspect-based sentiment analysis.

References

1. *Blei M., Ng A., Jordan M.* (2003), Latent Dirichlet allocation, Journal of Machine Learning Research, Vol. 3, pp. 993–1022.
2. *Brody S., Elhadad N.* (2010), An Unsupervised Aspect-Sentiment Model for Online Reviews, Proceedings of The 2010 Annual Conference of the North American Chapter of the ACL.
3. *Chetviorkin I., Braslavskiy P., Loukachevitch N.* (2012), Sentiment Analysis Track at ROMIP 2011, Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue", No. 11(18), pp. 739–746.
4. *Glorot X., Bordes A., Bengio Y.* (2011), Domain adaptation for large-scale sentiment classification: A deep learning approach, ICML, pp. 513–520.
5. *Hu M., Liu B.* (2004), Mining and summarizing customer reviews, International Conference on Knowledge Discovery and Data Mining (ICDM).
6. *Jakob N., Gurevych I.* (2010), Extracting Opinion Targets in a Single-and Cross-Domain Setting with Conditional Random Fields, Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2010).
7. *Jin W., Ho H.* (2009), A novel lexicalized HMM-based learning framework for web opinion mining, Proceedings of International Conference on Machine Learning (ICML-2009).
8. *Klekovkina M. V., Kotelnikov E. V.* (2012), The automatic sentiment text classification method based on emotional vocabulary [Metod avtomaticheskoj klassifikatsii tekstov po tonalnosti osnovannyj na slovare èmotsionalnoj leksiki], Digital libraries: advanced methods and technologies, digital collections (RCDL-2012) [Èlektronnye biblioteki: perspektivnye metody i tehnologii, èlektronnye kollektsii], Pereslavl-Zalessky, pp. 118–123.

9. *Li F., Huang M., Zhu X.* (2010), Sentiment analysis with global topics and local dependency, Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-2010).
10. *Liu B.* (2012), Sentiment analysis and opinion mining, Morgan & Claypool Publishers.
11. *Mikolov T., Sutskever I., Chen K., Corrado G., Dean J.* (2013), Distributed Representations of Words and Phrases and their Compositionality, Proceedings of NIPS.
12. *Mukherjee A., Liu B.* (2012), Aspect Extraction through Semi-Supervised Modeling, Proceedings of 50th Annual Meeting of Association for Computational Linguistics (ACL-2012).
13. Mystem, available at: http://company.yandex.ru/technology/mystem.
14. *Padró L., Stanilovsky E.* (2012), FreeLing 3.0: Towards Wider Multilinguality, Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA.
15. *Popescu A., Etzioni O.* (2005), Extracting product features and opinions from reviews, Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP).
16. *Řehůřek R., Sojka P.* (2010), Software Framework for Topic Modelling with Large Corpora, Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50.
17. *Rumelhart D., Hintont G., Williams R.* (1986), Learning representations by back-propagating errors, Nature, pp. 533–536.
18. *Socher R., Lin C., Ng A., Manning C.* (2011), Parsing natural scenes and natural language with recursive neural networks, Proceedings of the 26th International Conference on Machine Learning (ICML).
19. *Titov I., McDonald R.* (2008), A joint model of text and aspect ratings for sentiment summarization, Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2008).
20. *Turney P.* (2002), Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, Proceedings of the Association for Computational Linguistics (ACL), pp. 417–424.
21. *Turney P.* (2013), Distributional semantics beyond words: Supervised learning of analogy and paraphrase, Transactions of the Association for Computational Linguistics (TACL), pp. 353–366.
22. *Van der Maaten L., Hinton G.* (2008), Visualizing High-Dimensional Data Using t-SNE, Journal of Machine Learning Research, pp. 2579–2605.
23. *Zagibalov T., Carroll J.* (2008), Automatic seed word selection for unsupervised sentiment classification of chinese text, Proceedings of the 22nd International Conference on Computational Linguistics, Morristown, pp. 1073–1080.

ОБ ОДНОЙ ИЗ САМЫХ ЧАСТЫХ ЕДИНИЦ РУССКОЙ СПОНТАННОЙ РЕЧИ: *БЛИН* С ЛИНГВИСТИЧЕСКОЙ И СОЦИОЛИНГВИСТИЧЕСКОЙ ТОЧЕК ЗРЕНИЯ[1]

Богданова-Бегларян Н. В. (nvbogdanova_2005@mail.ru)

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Ключевые слова: русская спонтанная речь, функциональные единицы речи, Звуковой корпус русского языка, клитическое употребление

ONE OF THE MOST FREQUENT ITEMS IN RUSSIAN SPONTANEOUS SPEECH: *БЛИН* FROM LINGUISTIC AND SOCIOLINGUISTIC POINTS OF VIEW

Bogdanova-Beglarian N. V. (nvbogdanova_2005@mail.ru)

Saint Petersburg State University, St. Petersburg, Russia

The paper is dedicated to some peculiar functions of one of the most frequent items in Russian spontaneous speech, “блин”, which formally being a word, is in fact more of a functional item). Using a Russian speech corpus (‘One Speech Day’ sub-corpus) we explored the historical change of the item; from an interjectionally used euphemism for an extremely rude slang word meaning ‘whore’—through an acceptable colloquialism—to an almost meaningless clitic. So the evolution of this word begins at the point of being absolutely unacceptable in everyday speech, continues through being common and existing in any kind of neutral speaking, and ends as an ornamental word that probably lost the connection with its first meaning completely.
The final item does not have any meaning, lacks grammar categories, is not marked by intonation and has almost no emotional connotation. Normally such words are mostly used by men; but in this particular case gender does not play any role.

Key words: Russian spontaneous speech, functional speech items, Russian speech corpus, clitic

---

[1] Исследование выполнено при поддержке гранта РГНФ № 13-04-12022 «Информационная словарная система „Язык мегаполиса“».

Поворот лингвистического интереса от письменной речи, надежно зафиксированной в словарях и грамматиках, подчиненной узаконенным кодификацией правилам, — к речи устной, крайне нестабильной и часто далекой от устоявшихся представлений о нормативном и правильном, буквально ускользающей от исследователя (и потому усиливающей желание «поймать», зафиксировать и всмотреться), не только позволяет увидеть эволюцию языка фактически в ее реальном протекании, но и ставит перед лингвистами много новых вопросов и практического, и теоретического свойства. Уже не раз писали и говорили о специфике *речевой грамматики* и *речевого лексикона*, обсуждается вопрос о пересмотре старого или даже создании нового терминологического аппарата (*метаязыка*) для анализа речевых явлений, актуальной стала задача лексикографического описания всех разновидностей функциональных единиц устной речи — как речевых, так и условно-речевых, частотность которых в реальной речевой практике вынуждает сделать их главным объектом лингвистических описаний. Ср. одно из авторитетных мнений на этот счет: «В свете корпусной идеологии совершенно по-новому предстают приоритеты лингвистической теории. Теоретическая лингвистика последних десятилетий затратила огромные усилия на анализ сложных синтаксических явлений. Однако с точки зрения корпусного подхода эта работа не всегда полезна, поскольку многие такие явления в речевой реальности не обнаруживаются или обнаруживаются крайне редко. В то же время исключительно частотные явления устной речи, такие как хезитации, речевые сбои, регуляторные дискурсивные маркеры, парцелляции и т. д., практически не замечены лингвистической теорией» (*Рассказы о сновидениях* 2009: 27). Вслед за А. А. Кибриком и В. И. Подлесской, авторами этой пространной цитаты, свою задачу мы видим в том, чтобы «исправить этот крен и расширить эмпирическую базу лингвистического анализа» (*там же*).

Так, отдельного лингвистического внимания и детального анализа заслуживает, как представляется, буквально каждое из *слов-паразитов*, коммуникативные функции которых в речи не всегда поддаются однозначной трактовке («значимое» и «незначимое», «паразитическое», употребление прагматических маркеров «не всегда легко разграничить» — *Сиротинина* 1974: 71; см. также: *Шмелев* 2005: 519), каждый из *вербальных хезитативов* — во всем многообразии выполняемых им функций, и даже каждая выделенная коммуникативная функция той или иной дискурсивной единицы — во всем многообразии способов ее речевого воплощения (см., например: *Богданова-Бегларян* 2013).

В настоящем исследовании объектом такого пристального внимания стало «слово» *блин*, прошедшее за короткое время и буквально на наших глазах путь от бранного слова (детского ругательства), эвфемистического заменителя грубой инвективы, до просторечного, но уже привычного и очень распространенного междометия. Анализ материалов Звукового корпуса русского языка «Один речевой день» (ОРД) (см. о нем подробнее: *Звуковой корпус*… 2013, 2014) позволяет увидеть, что употребления данной единицы в нашей речи теряют уже и междометную природу (с передачей тех или иных эмоций), превращаясь в нечто орнаментально-клитическое, трудноопределимое не только в семантическом и грамматическом, но и в прагматическом аспекте.

Как бы ни относиться к явной экспансии этой единицы в нашей повседневной устной речи, ее функционирование однозначно заслуживает более пристального внимания, ср.: «от современной разговорной речи в ее нейтральном слое невозможно (со стилистической точки зрения) отсечь обширный репертуар нелитературных и окололитературных — сниженно-обиходных, просторечно-профессиональных, жаргонных и полужаргонных средств» (*Винокур* 1988: 54).

Нельзя сбросить со счетов и того факта, что в верхушке частотного словника ОРД (из расшифровок объемом более 350 тыс. единиц; максимально естественная, повседневная, речь носителей русского языка, записанная, что называется, «с диктофоном на шее») «слово» *блин* опережает практически все знаменательные слова. Первый полноценный глагол в этом словнике — *знаю* (608 употреблений; 40-е место; 0,30% от всего массива употреблений корпуса)[2]. При этом ни одного полноценного существительного в списке 150 самых частых единиц не обнаружилось вовсе. На 85-ом месте оказалось наше *блин* (305 употреблений; 0,15%), сопоставимы с ним по частоте *бл… дь* — 206 употреблений (116-ое место; 0,10%); *типа* — 199 (118,5 место; 0,10%); и *время* — 197 (121-ое место; 0,10%). Последние две единицы, как показывает анализ их контекстов, тоже по преимуществу употребляются не как имена существительные, а как различные дискурсивные слова (или компоненты дискурсивных единиц) (см. подробнее: *Звуковой корпус…* 2014).

«Слово» *блин* в рассматриваемом значении зафиксировано по преимуществу в различных словарях неформальной лексики, хотя и не только: см., например, «Большой толковый словарь русского языка» под ред. С. А. Кузнецова (*БТС* 2009) или «Словарь современного русского города» (*Словарь современного русского города* 1993).

Большинство словарей определяют *блин* как единицу *в знач. межд., вводн. Жарг. эвфем.* со значением выражения любых эмоций: досады, раздражения, удивления, иногда — восхищения, одобрения или восторга. Отмечается, что это «каламбурное употребление нейтрального слова вместо сходного по звучанию БЛ…ДЬ» (*Химик* 2004: 48; см. также: *Словарь современного русского города* 1993: 34; *Елистратов* 2002).

В Словоборге (slovoborg.su) также находим определение *блина* как слова-паразита в функции междометия; здесь дается и его оценка пользователями Интернета: 21 — ЗА, 12 — ПРОТИВ. Из такого соотношения оценок видно, что носители языка в большей степени склонны принимать эту единицу в указанной функции, чем отвергать ее. Ср. характерный стишок с просторов Интернета:

---

[2] Высокая частота в ОРД отрицательной частицы не (2-ое место в частотном словнике; 4924 употребления; 2,4% от общего количества единиц) позволяет предположить, что значительное количество употреблений этого знаю — из конструкции (я) не знаю, которая, по нашим данным (см. Звуковой корпус… 2014), почти в половине своих контекстов (48,7%) выступает как вербальный хезитатив, опережая по частотности все другие свои функциональные разновидности.

Это слово пришло из былин
И оно даже детям знакомо:
*Говорим мы привычное «**блин**»*
На работе, в дороге и дома.
Без него и беседа пуста,
*Скажешь «**блин**» — всё наладится быстро.*
Слово это у всех на устах:
И у дворника, и у министра.
На скамеечке — он и она,
Дышит нежностью каждое слово:
*— Для меня ты, **блин**, Машка, одна…*
*— Ах, как, **блин**, я люблю тебя, Вова…*

Словарь молодежного сленга уточняет значение этого междометия: ‘возглас выражения отрицательных эмоций, досады; ругательство’ (teenslang.su), из чего видно, что возможность выражения этим словом положительных эмоций, предусмотренная рядом других словарей, здесь не поддерживается. То же видим и в электронном русско-английском справочнике по разговорной речи (englishtown.com/speaking-english), который дает соответствующие варианты перевода этого слова на английский разговорный: *блин!* (также: *черт*! *Черт побери! Твою мать!*) — *lynchpin*, *Bugger! Pants!* — тоже исключительно в отрицательных, ругательных вариантах.

Именно такое — междометное (непременно эмоциональное) — употребление «слова» *блин* можно видеть и в письменных текстах, передающих особенности разговорной речи персонажей, ср. примеры из основного подкорпуса Национального корпуса русского языка (НКРЯ):

- *Да, **блин**, дела, подумали Ваня и капитан Медведев* [В условиях реального времени (2002) // «Культура», 2002.03.25];

- *— **Блин**, как он меня достал, — со стоном проговорила Елена Николаевна* [А. Геласимов. Фокс Малдер похож на свинью (2001)];

- *Это я? Во **блин**! — Это ты* [Л. Петрушевская. Маленькая волшебница // «Октябрь», 1996].

Орфография данных контекстов убеждает, что перед нами именно междометия, передающие те или иные эмоции говорящего.

Между тем, анализ материала ОРД (естественная устная повседневная речь самых разных носителей русского языка) показывает, что междометное употребление единицы *блин* не только не единственный, но и далеко не самый распространенный вариант ее функционирования — на его долю приходится не более 34,5% всех подобных словоупотреблений. Существенно чаще (65,5%) *блин* используется в нашей устной речи как своеобразная клитика: без лишней эмоциональной окраски, без всякого интонационного выделения, свойственного

междометиям, и зачастую без какой бы то ни было эвфемизации[3]. В подавляющем большинстве случаев (76,4%) эта единица занимает в высказывании позицию ***энклитики*** (постпозицию по отношению к знаменательному слову, некоторое предпаузальное усиление синтагматического членения), ср.:

(1) *кого это мне напоминает* ***блин*** */ я не знаю* (И1_ж33#К1_м?)[4]*;*

(2) *не / просто уже так привыкли* ***блин*** (И8_ж16#К2_м20)*;*

(3) *я раньше жила рядом* () *совсем рядом / с этим районом // *П я как узнала / что там* () *будет строиться / *П меня это ещё подстегнуло / чтоб уехал* ***блин*** (И9_ж27#К2_ж35)*;*

(4) *нет / просто / *П мне очень интересно / *П кто* () *ху из ху / какого происхождения потому что / *П какой-нибудь там не знаю немец / у него плохая погода* ***блин*** */ ну везде плохая погода* (И11_ж28#К1_м?)*;*

(5) *я вот вообще бл*[…]*дь*[5] *машину не хочу / вот серьёзно тебе говорю // после … после той … той аварии* ***блин*** */ я понял* ***блин*** */ что рано мне ещё на машине ездить* (И21_м27#К1_м?).

Из примеров видно, что *блин* появляется в речи и мужчин, и женщин, в том числе в их общении и с мужчинами, и с женщинами, преимущественно молодых людей (до 35 лет), часто эмоционально совсем не нагружен и прекрасно соседствует со своим нецензурным «прототипом» — см. контекст (6), — что плохо согласуется с представлением о его эвфемистической роли.

Существенно реже *блин* выступает в позиции ***проклитики*** (в препозиции по отношению к знаменательному слову, некоторое постпаузальное усиление синтагматического членения) (2,7%) или своеобразной «***интерклитики***» (в интерпозиции в синтагме, никак интонационно не выделено, акцентологически не прикреплено однозначно ни к предшествующему, ни к последующему слову) (20,9%):

---

[3] Такое, во многом клитическое, употребление в разговорной речи нецензурной лексики не раз отмечалось исследователями, ср.: «в этой функции нецензурные слова произносятся вставочно, для „связки слов“, они не выделяются в потоке речи интонационно, не выделяются громкостью, фонетически примыкают к предыдущему или последующему слову, используются фактически безоценочно и в известной степени — орнаментально, для придания речи эмоциональности» (*Стернин* 1999: 182–183). Думается все же, что клитические употребления «слова» *блин* утрачивают в современной речи уже и всякий элемент эмоциональности.

[4] Все примеры в данном разделе атрибутированы, с указанием номера информанта (И) и его коммуниканта (К), а также пола и возраста обоих. Данная реплика — из речи женщины 33 лет, обращенной к мужчине примерно такого же возраста (точных данных нет).

[5] В отличие от обычных в спонтанной речи запинок или обрывов слов, обозначаемых с помощью многоточия, таким образом ([…]) оформляются в примерах нецензурные выражения, в которых сознательно опущены некоторые буквы.

(6) *ну вот // *П да // мне тоже было бы жалко если* () *тем более они приедут такие наглые / **блин** как вот эти айзеры* (И9_ж27#К1_ж58#К2_ж35)*;*

(7) *слушай / *П в городе / *П **блин** я ду…* (э…э) */ Кама$ *Н вообще / *П супер полезно бл*[…]*дь // @ я по-любому блин* (И36_м40#К1_м40)*;*

(8) *я когда выходной / я все **блин** делаю / все наготовлю* (И10_м28#К1_ж?)*;*

(9) *вот та самая / любовь любовь которая / *П кто-то там // *П типа **блин** там / чувства и так далее // *П там всё то же самое / понимаешь ?* (И11_ж28#К1_м?)*;*

(10) *выключу я наверно свой телефон на всякий случай // *П что-то **блин** неспокойно у меня на душе вообще* (И21_м27#К1_м?)*;*

(11) *на х*[…]*й в цвет стен ! *П белый /* (…) *# ну я с… сразу подумал / белый я говорю / ему с… его спрашиваю / белый **блин** красить ?* (И36_м40#К2_м24).

Снова хорошо видно, что *блин* появляется в речи практически любых информантов, в любом коммуникативном акте, в числе говорящих появились уже люди 40 и 58 лет, часто *блин* в их речи оказывается в непосредственном соседстве со своим «прототипом» (7) и с другими непечатными выражениями (11), а также порой встраивается в достаточно протяженную хезитационную конструкцию, сближаясь по функции с вербальными хезитативами: см. примеры (1), (7). Даже поверхностный анализ лексики в приведенных иллюстрациях показывает, что сплошь и рядом *блин* для говорящих — вполне нейтральное слово.

Гендерной «привязки» этой единицы в исследованном материале практически не прослеживается, хотя в речи мужчин она все же несколько преобладает, ср. данные по ОРД: 57,8% употреблений «слова» *блин* в целом у мужчин против 42,2% в речи женщин, в том числе в его междометном употреблении: 59,5% (М) vs. 40,5% (Ж). Что касается возраста, то, как и можно было бы предположить, в большей мере это «слово» свойственно речи людей молодых и средних лет.

Любопытно, что в письменных текстах (основной подкорпус НКРЯ) таких примеров немеждометного употребления *блина* пока не зафиксировано (впрочем, самый поздний контекст такого рода — только 2002 г.), что лишний раз убеждает, что клитическое употребление единицы *блин* стало активным лишь в устной речи и лишь в самое последнее время. Письменная речь (художественные тексты) просто не поспела еще за развитием устного дискурса, ср.: «исходя из примата спонтанной диалогической речи в антропогенезе, можно утверждать, что в организации художественных текстов нет ничего, чего не было бы в спонтанной речи» (*Мурзин, Штерн* 1991: 161).

Заключая этот небольшой анализ, хочется повторить достаточно очевидную вещь: живая устная речь богата и разнообразна, внимательное отношение к ее единицам и их поведению может смягчить привычный обывательский негативизм

в отношении многих чисто речевых явлений и увидеть за ними не только небрежность нашего говорения, не только пренебрежение к языку, но и языковую эволюцию. Действительно, мы привыкли уважать результаты языковых процессов, произошедших давно, а то, что происходит на наших глазах, предпочитаем, не задумываясь, считать ошибками и отклонениями от нормы, ср.: «Забавно только то, что акцентологические процессы, имевшие место много сотен лет тому назад, воспринимаются как предмет самой серьезной и уважаемой науки, тогда как происходящие у нас на глазах изменения места ударения — как грубые и вульгарные ошибки» (*Николаева* 2007: 466). Т. М. Николаева говорит здесь об ошибках ударения, но это вполне применимо и ко всем другим особенностям устной речи.

Думается, что словарная статья на «слово» *блин*, со всем разнообразием его значений и функций, реализующихся в повседневной устной коммуникации, должна занять свое место в соответствующем словаре русской устной повседневной (не только экспрессивной!) речи.

Литература

1. *Богданова-Бегларян Н. В.* Кто ищет — всегда ли найдет? (о поисковой функции вербальных хезитативов русской спонтанной речи) // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог» (2013) (Бекасово, 29 мая — 2 июня 2013 г.). Выпуск 12 (19). В двух томах. Том 1. Основная программа конференции / Гл. ред. В. П. Селегей. М., РГГУ, 2013. С. 125–136.
2. БТС — Большой толковый словарь русского языка [Электронный ресурс] / Под ред. С. А. Кузнецова. ГРАМОТА.РУ, 2009 // http://www.slovari.gramota.ru.
3. *Винокур Т. Г.* Устная речь и стилистические свойства высказывания // Разновидности городской устной речи. Сборник научных трудов / Ред. Д. Н. Шмелев, Е. А. Земская. М., Наука, 1988. С. 44–84.
4. *Елистратов В. С.* Словарь русского арго (материалы 1980–1990 гг.) [Электронный ресурс]. ГРАМОТА.РУ, 2002.
5. Звуковой корпус как материал для анализа русской речи. Коллективная монография. Часть 1. Чтение. Пересказ. Описание / Отв. ред. Н. В. Богданова-Бегларян. СПб., Филологический факультет СПбГУ, 2013. 532 с.
6. Звуковой корпус как материал для анализа русской речи. Коллективная монография. Часть 2. Теоретические и практические аспекты анализа. Том 2. Звуковой корпус как материал для новых лексикографических проектов / Отв. ред. Н. В. Богданова-Бегларян. СПб., Филологический факультет СПбГУ, 2014 (в печати).
7. *Мурзин Л. Н., Штерн А. С.* Текст и его восприятие. Свердловск, Урал. ун-т, 1991. 171 с.
8. *Николаева Т. М.* Грубые ошибки или назойливая языковая тенденция? // Язык в движении. К 70-летию Л. П. Крысина / Отв. ред. Е. А. Земская, М. Л. Каленчук. М., Языки славянской культуры, 2007. С. 466–470.
9. Рассказы о сновидениях. Корпусное исследование устного русского дискурса / Ред. А. А. Кибрик, В. И. Подлесская. М., Языки славянских культур, 2009. 736 с.

10. *Сиротинина О. Б.* Современная разговорная речь, ее особенности. М., Просвещение, 1974. 143 с.
11. Словарь современного русского города / Под ред. Б. И. Осипова. М., Русские словари. АСТ Астрель. Транзиткнига, 1993. 565 с.
12. *Стернин И. А.* Некоторые жанровые особенности мужского коммуникативного поведения // Жанры речи. 2. Сборник научных статей. Саратов, Гос. учебно-научный центр «Колледж», 1999. С. 178–185.
13. *Химик В. В.* Большой словарь русской разговорной экспрессивной речи. СПб., Норинт, 2004. 768 с.
14. *Шмелев А. Д.* «Показатели хезитации» в русской устной речи // Язык. Личность. Текст. Сборник статей к 70-летию Т. М. Николаевой / Отв. ред. В. Н. Топоров. М., Языки славянских культур, 2005. С. 518–530.

References

1. *Bogdanova-Beglarian, N. V.* (2013) Those Who Search, Do They Always Find? (about Retrieval Function of Hesitatives in Russian Spontaneous Speech) [Kto ishchet — vsegda li najd'ot? (o poiskovoj funkcyi verbal'nyx xezitativov russkoj spontannoj rechi)] Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii «Dialog 2013» (12/19) (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference «Dialog 2013»). Moskva, pp. 125–136.
2. *BTS* (2009) — Great Dictionary of the Russian Language / S. A. Kuznecov (ed.) [Electronic resource] [Bol'shoj tolkovyj slovar' russkogo jazyka / Pod red. S. A. Kuznecova] [Elektronnyj resurs] // http://www.slovari.gramota.ru.
3. *Chimik, V. V.* (2004) Great Dictionary of Expressive Russian Speech [Bol'shoj slovar' russkoj razgovornoj ekspressivnoj rechi]. Sankt-Peterburg, 768 pp.
4. *Dreamrecitals*. Corpus-based Research of Russian Speech Discourse (2009) [Rasskazy o snovidenijax. Korpusnoe issledovanie ustnogo russkogo diskursa]. Moskva, 736 pp.
5. *Elistratov, V. S.* Russian Argo Dictionary (materials of 1980–1990) [Slovar' russkogo argo (materially 1980–1990 gg.) [Elektronnyj resurs] // http://www.slovari.gramota.ru.
6. *Murzin, L. N., Shtern, A. S.* (1991) Text and its Perception [Tekst i ego vosprijatie]. Sverdlovsk, 171 pp.
7. *Nikolaeva, T. M.* (2007) Gross Mistake or an Obnoxious Tendency? [Grubye oshibki ili nazojlivaja jazykovaja tendencyja?] Jazyk v dvizhenii. K 70-letiju L. P. Krysina (Language in its development. To Prof. L. Krysin 70th anniversary). Moskva, pp. 466–470.
8. *Shmelev, A. D.* (2005) Features of Hesitation in Russian Speech [«Pokazateli xezitacyi» v russkoj ustnoj rechi] Jazyk. Lichnost'.Tekst. Sbornik statej k 70-letiju T. M. Nikolaevoj (Language. Personality. Text. Collectionofarticles. To Prof. T. Nikolaeva 70th anniversary). Moskva, pp. 518–530.

9. *Sirotinina, O. B.* (1974) Peculiar Features of Modern Spontaneous Speech [Sovremennaja razgovornaja rech', jejo osobennosti]. Moskva, 143 pp.
10. *Dictionary* of Modern Russian City (1993) [Slovar' sovremennogo russkogo goroda / Pod red. B. I. Osipova]. Moskva, 565 pp.
11. *Speech* Corpus as a Base for Analysis. Collective Monograph. Part 1. Reading. Retelling. Description (2013) [Zvukovoj korpus kak material dl'a analiza russkoj rechi. Kollektivnaja monografija. Chast' 1. Chtenie. Pereskaz. Opisanie]. Sankt-Peterburg, 532 pp.
12. *Speech* Corpus as a Base for Analysis. Collective Monograph. Part 2. Theory and Practice of Speech Analysis. Vol. 2. Speech Corpus as a Base for New Lexicographical Projects (2014) [Zvukovoj korpus kak material dl'a analiza russkoj rechi. Kollektivnaja monografija. Chast' 2. Teoretichskie i prakticheskie aspekty analiza. Tom 2. Zvukovoj korpus kak material dl'a novyx leksikograficheskix proektov]. Sankt-Peterburg (in print).
13. *Sternin, I. A.* (1999) Some Special Genres in Male Communication [Nekotorye zhanrovyje osobennosti muzhskogo kommunikativnogo povedenija] Zhanry rechi. 2. Sbornik nauchnyx statej (Speech genres. 2. Collection of articles). Saratov, pp. 178–185.
14. *Vinokur, T. G.* (1988) Speech and Stylistic Features of the Saying [Ustnaja rech' i stilisticheskie svojstva vyskazyvanija] Raznovidnosti gorodskoj ustnoj rechi. Sbornik nauchnyx trudov (Types of urban speech. Collection of research papers). Moskva, pp. 44–84.

ANAPHORA ANALYSIS BASED ON ABBYY COMPRENO LINGUISTIC TECHNOLOGIES

Bogdanov A. V. (abogdanov@abbyy.com),
Dzhumaev S. S. (sdzhumaev@abbyy.com),
Skorinkin D. A. (dskorinkin@abbyy.com),
Starostin A. S. (astarostin@abbyy.com)

ABBYY, Moscow, Russia

This paper presents an anaphora analysis system that was an entry for the Dialog 2014 anaphora analysis competition. The system is based on ABBYY Compreno linguistic technologies. For some of the tasks of this competition we used basic features of the Compreno technology, while others required building new rules and mechanisms or making adjustments to the existing ones. Below we briefly describe the mechanisms (both basic and new) that were used in our system for this competition.

Key words: anaphora resolution, coreference resolution, syntactic analysis, syntactic-semantic analysis

Introduction

The main task of the ABBYY Compreno system is to convert the input text into a semantic structure that is a tree where nodes are concepts and arcs are relations between these concepts. For details see [1] and [4].

At the early stage of the analysis process the structure of a sentence is represented as a syntactic tree. The syntactic analysis of the input text is complete, i.e. every item of the input text takes some syntactic slot of some parent.

Then the syntactic tree is augmented with non-tree links. While tree links encode syntactic dominance, non-tree links capture conjunction, pronominal anaphora, PRO control, and other non-local dependencies between nodes.

Further follows the transition from syntactic to semantic structure. During this process every parent-child arc in the tree is interpreted, and each node gets a semantic role related to its parent. The switch from syntactic slots to semantic roles is possible because each lexeme has a diathesis description—a list of correspondences between the syntactic slots that can connect to it and their semantic roles. During this transition the nodes that were bound with a non-tree link are replaced with their controllers. Let us consider an example:

(1a) Input text
Мальчик дал девочке свое яблоко.

(1b) Syntactic tree without non-tree links

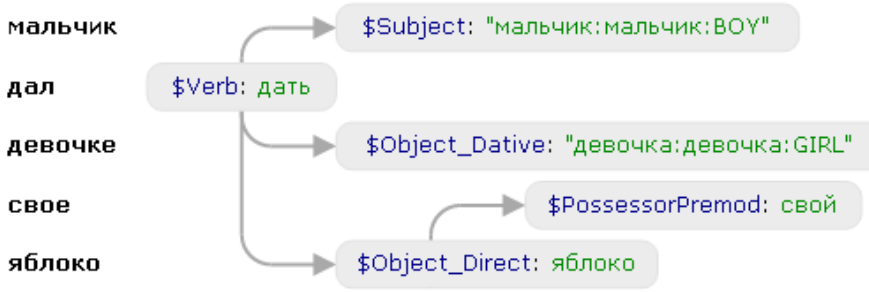


Fig. 1. Syntactic tree without non-tree links

(1c) Semantic tree with non-tree links

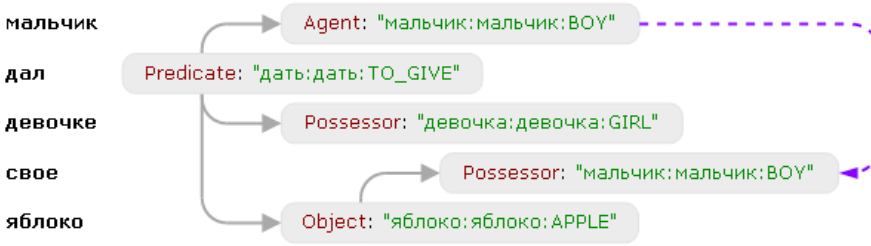


Fig. 2. Semantic tree with non-tree links

In (1c, fig. 2) the node *свое* is replaced with its non-tree controller *мальчик* which takes a semantic role of Possessor. If a controller or pronoun parent belonged to some other lexical class, its semantic role could be different. For example:

(2a) Input text
Мальчик знает своего врага.

(2b) Syntactic tree without non-tree links

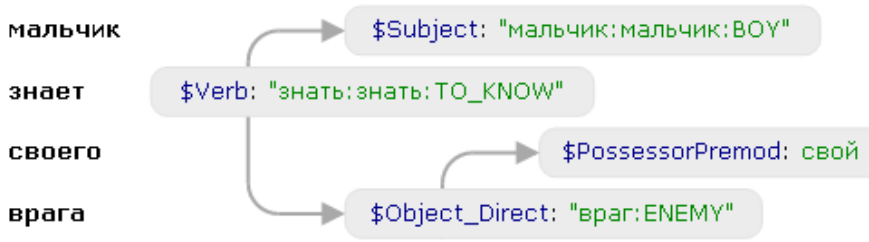


Fig. 3. Syntactic tree without non-tree links

(2c) Semantic tree with non-tree links

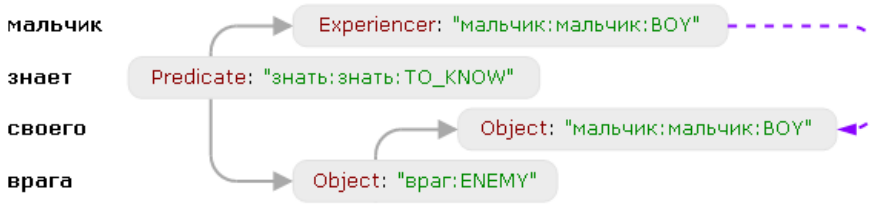


Fig. 4. Semantic tree with non-tree links

In (2) one can see that the same non-tree link as in (1) (between a Subject and a reflexive pronoun) results in a different semantic relation between the controlled node and its parent (semantic role of Object). This happens because when the controlled node is replaced with its controller, the semantic role is chosen depending on the lexical classes of both the controller and the node's parent. If more than one semantic role is possible for a given pair of items all of the possibilities are estimated and the best of them is chosen.

This mechanism of choosing semantic roles for the controlled node also helps us choose the most convenient controller for a given node, as demonstrated below.

1. Anaphora

1.1. Pronominal anaphora

One of the types of non-tree links in the Compreno system is pronominal anaphora. Pronominal anaphora resolution is an existing feature of the system, and therefore we did not have to build any special mechanisms for the purposes of the competition.

The pronominal anaphora rules are triggered if the system finds certain pronouns in the input text. Among such pronouns are: *он, она, оно, они, я, мы, ты, вы, себя, свой, друг друга, таковой* and some others. Each pronominal anaphora rule consists of the following components:

(3)

- list of pronouns that trigger the rule
- description of possible paths (via syntactic slots) from a possible controller to a pronoun
- description of possible properties of a controller
- a rule of agreement between a controller and a pronoun
- linear direction of the link (whether controller is to the left of the pronoun or to the right)
- value of the link

For example in (1) the appropriate rule chooses the Subject node as a controller because there is no path from the Dative object in this rule and there is a path from the Subject.

A description of possible properties of a controller is used to exclude controllers that are obviously impossible, for example such non-referential noun phrases as *в 2014 году, в трактористы, с моей точки зрения, в одностороннем порядке, по его требованию* etc.

In unambiguous examples like *Мальчик любит девочку. Она красивая*. the appropriate rule will choose *девочку* as a controller due to the agreement rule which says that in this anaphora rule a controller must have the same gender and number as a pronoun.

Now let us take an ambiguous example:

(4a) Input text
Мальчик любит этот дом—он его строил.

At the early stage of the analysis process we have a syntactic tree as follows:

(4b) Syntactic tree without non-tree links

Fig. 5. Syntactic tree without non-tree links

Then, as the pronouns (*он, его*) are found in the text, the anaphora rules are triggered and produce following links:

(4c) link 1:
Proform "он"; ProformParent "строить"; ProformSlot Object_Direct; Controller "дом"
link 2:
Proform "он"; ProformParent "строить"; ProformSlot Object_Direct; Controller "мальчик"
link 3:
Proform "он"; ProformParent "строить"; ProformSlot Subject; Controller "мальчик"
link 4:
Proform "он"; ProformParent "строить"; ProformSlot Subject; Controller "дом"

Then all possible sets of the non-tree links are formed (in every set, for one pronoun there is no more than one controller, which means that a pronoun may not have

a controller) and for each set the system seeks to replace a pronoun with its controller and choose a semantic role for it. It gives us a set of possible syntactic structures with replaced pronouns. These structures are ranked depending on the semantic compatibilities of all the items in given semantic roles (for details on the semantic compatibility and its evaluation see [4]). The best structure is chosen as a result of the analysis.

(4d) Semantic tree with non-tree links

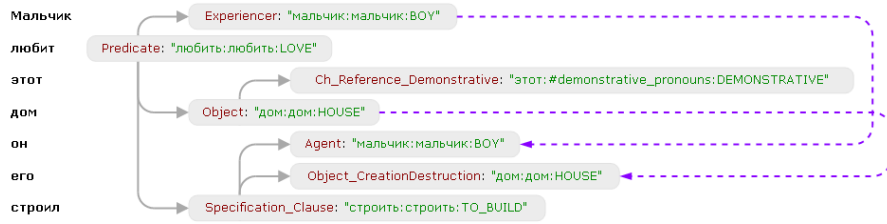


Fig. 6. Semantic tree with non-tree links

In (4d, fig.6) one can see that the pronoun *он* is replaced with its controller *мальчик,* which takes the semantic role of Agent. In its turn, the pronoun *его* is replaced with its controller *дом*, which takes the semantic role of Object_CreationDestruction.

That is how semantic compatibility between possible controllers and pronoun parents helps us in anaphora resolution.

1.2. Relative anaphora

Another type of non-tree links that is used in the Compreno system and was included in our competition links set is relative anaphora. By this term we mean a link between a noun phrase and a relative pronoun of a relative clause governed by this noun phrase like in example *Мальчик, который пришел.*

Links of this type are also drawn by special rules which have almost the same components as in (3) except that relative pronouns, unlike personal pronouns, must always be controlled, i.e. if for a given relative pronoun a controller is not found, then the whole structure is considered invalid. In semantic structure relative pronouns are also replaced with their controllers and choose appropriate semantic roles, which also helps choose the best controller among possible candidates relying on semantic compatibility.

Of course, a range of possible controllers in this case is much narrower than in the previous one, because a controller of a relative pronoun must govern its relative clause, and this information is stored in a corresponding rule as a description of possible paths between a controller and a pronoun. But even relative anaphora may have ambiguous cases, such as:

(5a) Input text
Мальчик видит игрушку девочки, которая пришла.

In (5a) for disambiguation the system should recognize that a girl is more likely to be able to walk than a toy. And this information can be obtained only from the semantic compatibility between a controller and pronoun parent. So for this sentence the semantic tree looks as follows:

(5b) Semantic tree with non-tree links

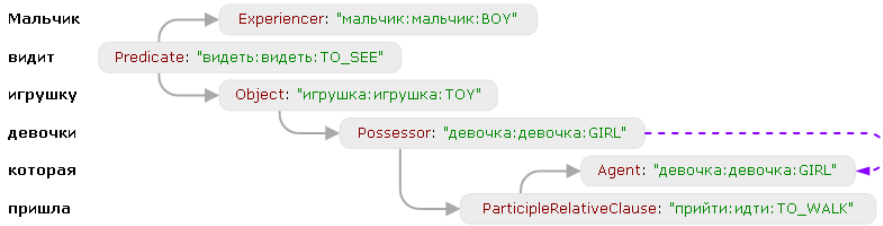


Fig. 7. Semantic tree with non-tree links

In (5b, fig. 7) relative pronoun *которая* is replaced with its controller *девочка,* which takes a semantic role of Agent. The structure where *игрушка* takes some semantic role of *прийти* was also considered, but was dismissed as having lower value.

2. Coreference

As challenging as it is, pronominal anaphora nevertheless represents only a limited subclass of the reference phenomena. Full-scale coreference resolution requires the ability to connect two separate nouns or noun phrases that refer to the same entity. The task gets especially complicated if the noun phrases in question have no string overlap at all, like *Obama* and *president* (compare *Barack Obama* and *Obama*, which is a relatively simple case)—a problem known as the 'opaque mentions' [3].

We regularly face this and other coreference-related issues in our ongoing work on named entity recognition (NER) and fact extraction. Relying on this experience, we are inclined to view coreference resolution as a subtask of entity recognition and identification in the broader sense of the word.

Even though the gold standard collection issued by the organizers did feature some examples of coreference between objects that could not be defined as named entities, these samples were relatively few. An overwhelming majority of coreferents tend to represent some kind of separate entity, either named or at least distinct and identifiable. Moreover, in most cases it was one of the 'big three' of NER—a person, a location or an organization. Therefore our approach mainly consisted in adjusting a set of ready-made entity extraction and identification rules to this particular task of coreference resolution. Nevertheless, some particular subtypes of coreference that could not be covered by the existing rules forced us to implement several new mechanisms, most notably a tool for graph-based semantic similarity measure that is described in the last section of this paper.

2.1. Candidate extraction

Two main stages of the process in our case are traditional for coreference resolution (see [2] for example), and include a) collecting all the probable candidates and b) filtering out those that do not seem to corefer with any other candidates. During the first stage we attempt to extract all the objects that could be identified as entities. Our entity extraction rules are generally based on the results of the ABBYY Compreno analysis and make use of the diverse linguistic information it provides (semantic classes, syntactic slots, semantic roles and many more, see [1] for details). The sets of rules vary for different types of entities. Here is a brief description of the core heuristics:

2.1.1. Person extraction

The task of person extraction in our system is subdivided into two major subtasks: detection of a person in a text and correct recognition of its attributes, i.e. name, surname, middle name and other parts of a proper name, if they are present. Extraction of attributes is essential for further identification of different textual instances as one person, as will be shown in the next sections.

The most obvious and straightforward way to locate a person in a text is by looking for a known personal proper name with capitalization. However, this simplistic approach alone rarely yields tolerable results, especially in terms of recall. First of all, even the most exhaustive databases cannot claim to have all the possible names and surnames, inevitably forcing a researcher to deal with the unknown ones. Secondly, there are many ambiguous names (*Bob, Virginia, Слава*), and even ones that lack ambiguity as such can still be used as proper names for entities other than human individuals (*пароход «Иван Федорович Крузенштерн», ресторан «Пушкин»*). Thirdly, a person can be referred to by a non-capitalized common noun/noun phrase (*мальчик, мужчина, космонавт, глава государства, state senator*).

The first problem—when a personal name is absent from the dictionary—can be addressed via syntactic and/or semantic structure. For instance, if a particular node of a parse tree has been labelled as an "UNKNOWN_BEING", we might try and look at the semantics of its parent. If the upper node turns out to be a name of a profession, a rank, an honorific or a nobiliary particle, chances are high that the node in question is a surname.

(6a) Input text
Я зашел к капитану Харгуду.

(6b) Semantic tree with syntactic slots

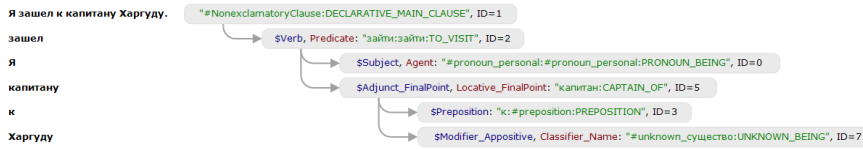


Fig. 8. Semantic tree with syntactic slots

Other personal markers include date of birth (*Гельмгольц, 1989 г.р.*), bracketed constructions with foreign words (*Кхиеу Порн (Khieu Porn) стал жертвой своих земляков*) or locations (*Вик Уайлд (Россия)—1 место*), certain verbs with strict selectional restrictions (*жениться, свататься*).

The issue of name ambiguity can be partially resolved by taking into account quotation marks and syntactic structure, which in some cases are determined by a particular meaning of an otherwise ambiguous proper name. Broader context might be helpful as well.

Addressing the third difficulty, when a referent of a person is a common name or a name phrase with no capitalization (*prime minister*), is particularly important for coreference resolution. However, at the stage of detection such cases pose little trouble—we basically mark any lexeme that fits semantically to define a person, is singular and complies with several other grammatical restrictions (so *prime minister* would fit and be extracted, as well as a *cosmonaut* or a *girl*).

As for the second subtask, the correct extraction of attributes (i.e. name parts) relies heavily on the common standards of writing down personal names. For instance, if we encounter a single initial followed by a capitalized word, the latter is usually a surname. Generally a complex personal name in our system is represented as a subtree with a surname or an initial as the top node. Its children might be a first name, several middle names or a patronymic, as well as initials or a part of a complex surname.

2.1.2. Organization & location extraction

The organization extraction rules fall into two main categories. Rules of the first category focus on keywords in the name of an organization itself and extract relatively straightforward mentions like *компания Тогрус* or *ОАО «Ромашка»* or *Cobham ltd*. They also deal with instances of enterprises and government bodies that are already known to the system by name.

Rules of the second category extract more obscurely-named organizations and rely on the context—mainly semantic classes and syntactic slots, but semantic roles are used sometimes as well. For instance, a rule that handles examples like *Он уволился из Омскэлектро* or *He resigned from RTRT* looks for a node with a semantic class "TO_RETIRE" and then creates an organization on its child provided that the latter has the semantic role of Locative_InitialPoint and is capitalized. Another rule that deals with corporate acquisitions (*Yahoo bought Tumblr*) requires a node with a semantic class "TO_ACQUIRE" with an Object in quotation marks among its children, while another child in the role of Possessor should not be a person (to exclude examples like *Vasya bought Sony Play Station*).

Proper names of the extracted organizations are stored as their 'identifier' attributes. Later on they are used at the identification stage.

Location extraction is based on the same principles. Keywords (*страна, город, озеро, bay, -city, creek* etc.) and sets of known proper names serve as the most reliable features, while previously unknown entities are derived with help of syntactic-semantic patterns. There are also additional stop-productions within the rules that do not allow the extraction of a known location in case it is used as a proper name for some other kind of named entities (*кафе Бомбей*).

The set of entities that are subject to extraction is not limited to these three types and includes a broad range of information objects from military aircraft to laws.

In these cases the general approach is quite similar to the one described above (while the exact properties of the extracted objects are, of course, different).

2.2. Identification and filtering

The first stage of the whole process can be described as a recall-oriented one, yielding a vast amount of referring expressions for further filtration. During the second stage the collected entities go through the identification process. The items identified as referring to one real-life object remain and form a coreference chain together, while the ones left without a pair are sifted out. This process determines the overall precision of the system, at the inevitable cost of decreasing the recall whenever an identification failure occurs. The identification rules rely chiefly on the attributes extracted during the entity extraction process. Following is a brief description of these rules for various entity types:

2.2.1. Person identification

The backbone of the identification of human-like entities is the intersection of attributes (name parts). For each pair of extracted persons the attributes are compared one by one, and if there is enough intersection and no contradictions, the objects can be merged. The discrepancy in gender prevents merging, so in case like *Иванов получил зарплату. Иванова рада* the entities will not be merged, whereas the two mentions of the same surname in *Иванов получил зарплату. Иванова обуяла радость* will be identified as relating to one person (this example demonstrated the advantages that complete syntactic-semantic analysis brings to coreference resolution).

Another way of person identification is via syntactic patterns combined with semantic restrictions. For instance, if a certain node with a person object attached to it has a nominal complement, we attach a special auxiliary link from the object to that complement. Then, if the same lexeme as in complement occurs elsewhere in the text, a second person is going to be extracted and the two person objects will merge due to that special link. Consider an example:

(7a) *Бьорндален—великий биатлонист. Спортсмен показал высший класс на олимпиаде в Сочи. Биатлониста такого уровня нельзя списывать со счетов и после 40 лет.*

In the first place our extraction rules locate three entities—*Бьорндален*, *биатлонист* and the second *биатлонист*. The two mentions of *биатлонист* are then merged into one person on the grounds of having similar semantic class, and after that the syntactic structure of the first sentence is used to identify *биатлонист* with the surname *Бьорндален*[1].

[1] Since the organizers of the contest chose not to consider coreference between a subject and its nominal complement, we did not connect them either. The described mechanism, nevertheless, was still used to identify and merge entities in the broader context. So in this particular case our coreference chain would show the connection between *Бьорндален* and *биатлонист* from the third sentence, but no visible link between the surname and the first *биатлонист* in the complement slot.

In order to extract the entire coreference chain from the last example one also has to identify биатлонист/Бьорндален with *спортсмен*. Fortunately, possession of an extensive semantic hierarchy allows us to do just that by incorporating certain WordNet-style graph-based metrics of semantic similarity into the identification process. In this particular case by traversing the hierarchical tree we find out that *спортсмен* is the direct hypernym of *биатлонист* and thus probably refers to the same person.

2.2.2. Organization and location identification

Organizations and locations are usually merged on the basis of their identifiers' (i.e. proper names) intersection. In addition to that there is a semantic similarity rule analogous to the one in person identification that was described above. Such a rule would merge *Роскосмос* and *контора* or *Роснефть* and компания in the following examples:

(8a) ***Роскосмос*** *запустил конкурента Google Maps. Государственная* ***контора*** *же, и деньгами налогоплательщиков работа оплачена.*

(9a) ***Роснефть*** *может получить контроль над всеми аэропортами Киргизии. Российская* ***компания*** *подписала меморандум о приобретении не менее 51 % ОАО «Международный аэропорт Манас».*

The identification will be possible because both *Роснефть* and *Роскосмос* are present in the semantic hierarchy and their semantic classes descend from these of the words *компания* and *контора*.

2.3. Adjustments for uncategorized entities

As has been mentioned before, the task of coreference resolution is not exactly limited to the identification of certain entities like individuals or organizations. In some cases coreferring expressions represent a real-life object that does not fall into any major entity category, and yet it is certainly supposed to be extracted.

A considerable share of such cases is constituted by demonstrative pronouns appearing as determiners (*лошадь—эта кляча; призрак—тот самый обозлившийся на него дух; аппарат—это устройство*). The resolution of this kind of coreference obviously requires some sort of semantic similarity data. As in case with common-noun persons, we use graph-based method. The idea behind this method is simple up-and-down tree traversal of the semantic hierarchy that yields synonyms as well as direct and indirect hypo/hypernyms. Whenever a demonstrative pronoun with a noun parent is encountered, the system launches a tree traversal procedure and the previous context is searched for a semantically similar noun. Here is an example from the test corpus of the competition:

(10a) *Я помню замечательный эпизод, когда она похвасталась нам с Володей Черняевым (он сейчас успешно работает в театре у Юрия Любимова) каким-то дорогим* ***одеколоном, который*** *она приобрела для молодого супруга. Мы попросили понюхать этот* ***парфюм.***

The semantic class of парфюм (“PERFUMES”, which also includes *парфюмерия*) is the direct ancestor of the semantic class of *одеколон* (“EAU-DE-COLOGNE”), which enables us to unite the two objects. The relative pronoun *который* is replaced by its controller *одеколон* and attached to the coreference chain as well.

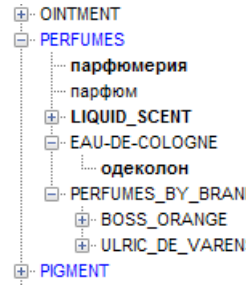


Fig. 9. A segment of the semantic hierarchy

Another example from the test corpus:

(11a) *Скоро ужасную* ***клячу****, словно сбежавшую с живодерни увидали и другие зрители. Люди смеялись, удивлялись, спрашивали, негодовали. Как могла попасть сюда эта* ***лошадь****?*

In this case two coreferents a) evidently represent an unnamed entity and b) are stylistic synonyms rather than hypo-hypernyms. In our semantic hierarchy the lexical classes *лошадь* and *кляча* exist within the same semantic class, and therefore the rule relying on demonstrative pronouns and semantic similarity applies to them as well.

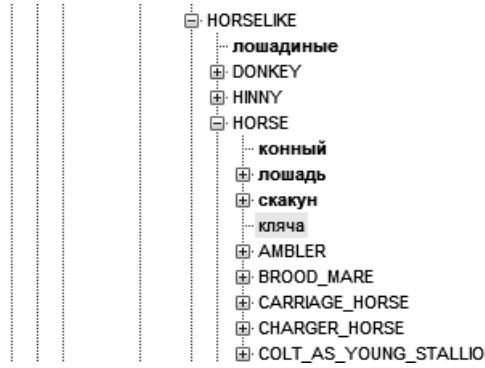


Fig. 10. A segment of the semantic hierarchy

Our experiments with the gold standard showed that this particular rule has very limited effect on the overall performance of the system, because the gain in recall is almost negated by the loss in precision, leaving F-measure increased by no more

than a few per mille. But that can be explained by inconsistencies in the corpus markup (many legitimate cases of coreference with demonstratives were left unmarked by the contest organizers) and relative scarcity of such cases in the provided texts.

Unfortunately, our attempts to use this sort of semantic similarity methods on a broader scope did not prove successful, yielding too many false positive hits. However, it is acknowledged that most of the attempts to detect such 'opaque mentions' (i.e. with no string overlapping of nouns) tend to decrease precision significantly more than improve recall [3].

Another crude recall-oriented adjustment is simply the extraction of all the nodes with capitalized lexemes (except for those in the beginning of a sentence, of course) as well as lexemes and expressions in quotes. Each of them received two identifiers, a lemma of a given lexeme and the original word form that appeared in the text. Thus an information object *Нацбест* in *лауреат Нацбеста* has two identifiers—normalized *Нацбест* and original *Нацбеста*, which in one case helped us to identify two coreferents despite the normalization failure. At the identification stage such candidates were compared to each other and merged in cases of identifiers matching. Of course this adjustment is limited to unknown entities only and does not apply to persons or organizations.

Conclusion

Our approach to anaphora and coreference resolution has an obvious bias towards deep linguistic analysis (rather than the use of statistics and machine learning) and can be described as rule- or model-based. Such approaches are known to be relatively labour-intensive and have their limitations. However, the use of deep semantic data allows our system to perform well in many challenging cases like ambiguous examples of pronominal anaphora or 'opaque mentions' of coreferring expressions. Linguistic information also enables us to avoid such typical false positives as individuals with similar surnames but different gender.

We evaluated our system's anaphora resolution on a part of the training corpus. Since there were some inconsistencies in the gold standard, we double-checked all the discrepancies manually, so that the result was not lowered by the correct pairs detected by the system but absent from the training markup. This semi-automatic evaluation showed the F-measure of 0,644. We chose not to evaluate coreference resolution ourselves due to lack of agreement on evaluation metrics in this particular field (since whole chains are supposed to be evaluated rather than just pairs). It is expected that by the time this paper is published the organizers will have revealed the results of the independent evaluation.

References

1. *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.
2. *Lee H., Peirsman Y., Chang A., Chambers N., Surdeanu M., Jurafsky D.* (2011), Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task, Proceedings of the CoNLL-2011 Shared Task, Portland, Oregon, USA, pp. 28–34.
3. *Recasens M., Can M., Jurafsky D.* (2013). Same Referent, Different Words: Unsupervised Mining of Opaque Coreferent Mentions, Proceedings of NAACL-HLT 2013, Atlanta, Georgia, USA, pp. 897–906.
4. *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.

ДИСКУРСИВНЫЕ СЛОВА И РЕФЕРЕНЦИЯ В ПРОЦЕССЕ ПОНИМАНИЯ СООБЩЕНИЯ

Борисова Е. Г. (efcomconf@list.ru)

Московский городской педагогический университет, Москва, Россия

В важнейшем для понимания сообщения процессе — установлении референциальных связей в тексте — определенную роль могут играть дискурсивные слова (частицы, вводные слова). Рассматриваются некоторые дискурсивные слова, которые в той или иной степени связаны с маркировкой референциального статуса существительных и схожих с референцией процессов понимания предикатов.

Ключевые слова: референция, денотативный статус, дискурсивные слова, частицы, понимание

THE DISCOURSE WORDS AND REFERENCE IN THE PROCESS OF UNDERSTANDING

Borisova E. G. (efcomconf@list.ru)

Moscow City Teachers' Training University, Moscow, Russia

The article addresses some aspects of the process of understanding, namely the reference of the components of utterances. The referential activity of the Hearer is regarded as a part of his actions in analyzing the sentences. These actions of the Hearer can be corrected by the Speaker with the help of discourse markers, modal particles etc. These entities are no markers of a referential status of nouns, still they can help reveal this status in some complicated cases, as follows:
A non-actualized (though definite) name is used as a topic of an utterance. The Russian particle—*to* that marks such topics can help reveal the definite status of the name. Some other complicated cases of topic formation can be marked by particles *vot* and *von*.
The word that denotes something known (maybe mentioned in the previous context) can be revealed with the help of the particles *vot*, *imenno*, *kak raz*. It concerns not only nouns but also predicates.
The indefinite status of a noun can be demonstrated with the help of the particle *tam*, which is used to denote unimportance of some fact or a noun.

Key-words: reference, referential status, discourse words, modal particles, understanding

1. Референция как часть понимания сообщения

К референции обычно относят средства соотнесения имен с объектами описываемого мира. (Падучева 1985, 7) Практически все определения говорят о реальном мире, что породило немало парадоксов и оговорок. Однако и лингвистика, и логика добились немалых результатов в обходе подводных камней, связанных с реальностью, тем, что стали опираться на «возможные миры», в частности, на представление реальности в дискурсе. Поэтому и (Падучева 1985), и более поздние работы ориентируются на «возможные миры» как отражение содержания сообщения.

Возникшее в рамках формальной логики понятие стало использовать в лингвистике для описания таких явлений, как категория определенности в артиклевых языках, неопределенные местоимения во многих языках, включая русский, а также некоторые вопросы коммуникативной организации, порядка слов и т. п.

Период расцвета изучения референции падает на 1980ые годы, когда было выполнено несколько глубоких и оригинальных исследований о возможных способах отражения референциального статуса имени в безартиклевом русском языке. В дальнейшем исследования приобрели более практический характер, т. к. вопросы определения статуса при анализе текста и выборе способов его выражения имеют прикладной характер, важный и в аспекте компьютерной обработки текста, и в вопросах преподавания иностранных языков, редактирования, перевода и др.

В большинстве моделей и описаний процессов анализа (понимания) текста установление референций составляет важный этап.

Примерно в те же годы происходило и развитие дискурсивного и когнитивного подхода к языку, в рамках которого право на существование получило представление о речемыслительной деятельности человека. Это позволило ввести деятельностный аспект в процесс установления референции. В исследованиях появляются попытки не просто описать какие-то признаки языковых единиц и систем, а сделать это с учетом возможных действий участников общения. Этот подход связан с развитием прагматической парадигмы в языкознании, однако его корни уходят вглубь к динамическим моделям типа модели «Смысл-Текст».

Одним из первых обращений к деятельности участников общения следует отметить ранние попытки представить различия между неопределенными местоимениями *any*, *some* в английском и *-то*, *-нибудь* в русском как различия в деятельности человека отражающего какую-то реальность. В частности, *-то* предполагает, что выбор уже осуществлен, а *-нибудь* — что это еще впереди. (В отечественной традиции начало этому положили О. Н. Селиверстова, И. М. Кобозева и др.).

Деятельностный подход затронул широкие слои языковой системы. В частности, в нем был по-новому представлен взгляд на дискурсивные слова. Поэтому обращение к дискурсивным словам в свете проблем установления референциального статуса имени существительного представляется закономерным и перспективным.

2. Дискурсивные слова в речемыслительной деятельности

В деятельностной (прагматической) парадигме языкознания понятие «дискурсивные слова» стало вытеснять ориентированное на формальные понятия «частица» (Modalpartikeln, Modal particles). С одной стороны, частицы обладают слишком различными функциями. Например, наряду с усилительными и выделительными частицами рассматриваются и формообразующие частицы, явно не являющиеся отдельными лексемами. А с другой стороны, схожие функции маркировки особенностей дискурса могут выполнять и другие части речи, например, наречия и фраземы в роли вводных слов: *собственно, честно говоря, фактически*.

В 1975 году был предложен термин Pragmalexeme т. е. «прагматические лексемы» (автор Р. Ратмайр), а затем, пожалуй, в еще более широком смысле Discourse markers (Д. Шиффрин), что и было переведено на русский как «дискурсивные слова». Мы не останавливаемся здесь на возникающих время от времени дискуссиях о соотношении всех этих терминов. Для данной работы важно, что в одном понятии»дискурсивные» объединены языковые средства со сходной прагматической функцией, которая и является основным предметом рассмотрения.

Дискурсивные слова выполняют разнообразные функции в тексте, привлекая особое внимание к объектам или событиям (а иногда и отвлекая), управляя возможными импликатурами. Нередко на основании дискурсивных функций формируются и синтаксические — указание на причинные связи (частицы *же*, *-то, ведь*), уступительность (*-то*) и т. п.

Рассмотрим основные случаи употребления дискурсивных слов, которые так или иначе могут влиять на восстановление референтных связей в процессе понимания.

Процесс референции в деятельностной модели языка представляет собой (огрублено) поиск референта — объекта в картине мира слушающего, который является предметом сообщения. В случае его (их) нахождения, мы говорим об определенной референции. В тех случаях, когда объекта еще нет в картине мира слушающего, осуществляется неопределенная референция, которая позволяет ввести новый объект и далее говорить о нем как об определенном (интродуктивная функция):

(1) *Вечером …года по одной из окраинных улиц шел студент* (неопределенная референция)*, одетый бедно, но с достоинством. Под мышкой у студента* (определенная референция) *был сверток.*

или просто упомянуть в качестве примера и т. п.

Обычно, как в диалоге, так и в нарративе, определенная референция осуществляется благодаря активизации того фрагмента картины мира, куда попадает референт. Самый тривиальный случай — это упоминание объекта, недавно названного в том же тексте, как в примере (1). Нередки и воспоминания о неупомянутых объектах, но известных и связанных с описываемым фрагментом — в случае, если они однозначно устанавливаются в контексте. Однако, как показали исследования, немало таких случаев, когда у слушающего могут возникнуть сомнения в референции из-за того, что он неправильно понял, о ком идет речь.

Дискурсивные слова имеют функцию, отличающуюся от функции указательных частиц, местоимений и других средств указания на определенный объект или на отсутствие такого объекта. (Мы не исключаем из рассмотрения усилительные частицы, образованные из указательных слов и местоимений, которые будут рассмотрены ниже. Однако в дискурсивной функции они имеют свои особенности). Частицы и вводные слова управляют вниманием слушающего, воспринимающего связную речь. И тут на долю частиц приходится управление вниманием, что часто выливается в маркировку известности объекта или события или, напротив, его неожиданности в данном контексте, важности для говорящего. В каких-то ситуациях такая неожиданность (или, напротив, заведомая известность) могут влиять на действия слушающего, ищущего референт для имени или местоимения.

Нами выявлено три таких случая. Но в целом влияние дискурсивных слов на процессы понимания и, в частности, установления референции шире. Однако иногда это влияние недостаточно значительно, носит эпизодический характер и, возможно, не заслуживает того, чтобы его принимали во внимание. Не исключено, что более тщательное изучение проблемы позволит обнаружить какие-то иные механизмы влияния на процессы референции, в которые так или иначе могут быть вовлечены дискурсивные слова. Пока же рассматривается использование частиц в ситуации сложного поиска денотата, при подтверждении определенного статуса повторяющихся имен и в ситуации задания неопределенного референциального статуса. Кроме того, рассматриваются процессы, близкие к референции, однако относящиеся к предикатам (глаголам, прилагательным).

3. Дискурсивные слова в сложном поиске денотата

Обращение к объекту с определенной референцией наиболее тривиально осуществляется в том случае, если он активизирован в тексте: уже упоминался или легко выводится из предыдущего контекста. Однако нередко в диалоге или монологической речи бывают случаи, когда вполне известный объект, даже называемый именем собственным, не активизирован. И определенность этого объекта для говорящего не превращается в определенность для слушающего.

Рассмотрим предложение

(2) *Профессор выпустил монографию.*

Для слушающего в данном контексте возможно понимание «Некий профессор (из числа, конечно, знакомых, но неопределенный) издал книгу». Оно было бы вполне естественно в беседе с предыдущей репликой:

(2’) *Что у них на кафедре, как всегда? — Профессор выпустил монографию, доценты разработали методички.*

Однако более вероятным будет понимание как указание на какого-то определенного профессора, который может быть отождествлен и говорящим, и слушающим. В этой ситуации слушающий может испытать затруднения при идентификации и, если ему не хватает указаний, может запросить дополнительную информацию:

(2'') *Это который?*

Если говорящий убежден, что оба участника могут иметь в виду одного и того же человека, он может употребить частицу *-то*:

(2''') *Профессор-то выпустил монографию.*

В таком случае может иметься в виду только определенный, известный обоим собеседникам человек.

Аналогичные функции могут выполняться частицей *-то* и при имени собственном. Считается, что имя собственное это способ определенной референции. Однако реально имя собственное может обозначать большое множество людей, и даже среди общих знакомых его могут носить несколько человек:

(3) *Лена Борисова в Фейсбуке разместила фото.*
— А она разве фотографирует?
— Да это другая. Две Елены Борисовы в одной группе. — Перебор!

Поэтому во фразе

(3') *Борисова**-то** премию получила!*

Частица *-то* выполняет ту же функцию: актуализует неактуализованный ранее определенный объект.

Частица *-то*, как отмечено в (Борисова 1989), используется в тех случаях, когда необходимо привлечь дополнительное внимание к теме высказывания, например, при противопоставлении:

(4) *Семья-**то** большая, да два человека всего*
мужиков-то (Н. А. Некрасов Крестьянские дети).

В рассмотренном нами случае с примером (1) частица *-то* маркирует привлечение внимания к неактуализованной теме (Борисова 1982). В рамках деятельностного подхода результатом такого привлечения внимания должна стать актуализация, то есть обнаружение объекта, который может иметься в виду. В этом случае определенный статус объекта не будет вызывать сомнений.

Таким образом, если некоторое имя не упоминалось в данном тексте, оно сопровождается частицей *-то*, которая не имеет ни противопоставительного, ни причинного значения, можно считать с большой долей вероятности, что перед нами определенная референция. Исключения возможны для родовых обозначений:

(5) *Корюшка-то уже берёт!* (пример записан автором)

Заметим, что частица *-то* очевидным образом имеет этимологическую связь с указательным местоимением *тот*. Именно это местоимение используется для формирования определенного артикля в славянском артиклевом языке — болгарском, где у артикля не только те же корни, но и такая же постпозитивная позиция: *любовта* ‘любовь’. Это приводит к тому, что в некоторых работах утверждается, что частица *–то* в русском литературном языке тоже может служить для выполнения артиклевой функции. Но надо иметь в виду, что использование частицы *-то* для указания на определенность встречается нечасто, частица может употребляться не только с именем существительным, но и с прилагательным, глаголом, наречием. Да и в случае выражения определенности это явление далеко от грамматикализации, хотя и может использоваться в процедурах установления референта.

4. Дискурсивные слова и отождествление в тексте

4.1. Еще один важный фактор, связанный с референцией, заключается в проблеме идентификации денотата в тексте. В артиклевых языках первое упоминание объекта маркируется неопределенным артиклем, последующие — определенным, что заставляет искать денотат для этого имени в числе актуализированных или относящихся к общим знаниям говорящего и слушающего.

В безартиклевых языках, к которым относится и русский, именование объекта не всегда дает основание для выяснения: это первое упоминание неизвестного объекта или называется уже актуализованный или известный заранее денотат ср.:

(6) *Я это еще с пятикурсниками обсуждала. Очень интересный был результат у одного. Он мне привел выкладки. Я было не поверила, но студент мне сказал…*

В этом случае для того, чтобы правильно понять, является ли студент тем, кто получил интересный результат, или это еще один из участников дискуссии пятикурсник, нужно использовать указательные или личные местоимения (*он сказал*, *этот студент сказал*) или обратиться к другим средствам указания на отождествление. В данном примере (записанном автором) была использована частица *вот*:

(6’) *… Я было не поверила, но студент* ***вот*** *мне сказал…*

Опознанию объекта как известного, имеющего определенный статус, способствуют частицы *вот, вон, как раз, именно*, а также уже упоминавшаяся частица *-то*. Их основные значения так или иначе связаны с установлением соответствия между наименованиями объектов или событий и объектами (событиями, лицами и т. п.):

(7) *«Если с драматургической точки зрения мелодрама предполагает возможность легко предвидеть развитие интриги, то» Огни рампы «представляют* ***как раз*** *тот фильм, в котором происходящее никогда не соответствует в точности ожидаемому»* (Андре Базен, Величие «Огней рампы» в книге: Что такое кино?, М., 1972, с. 211). [Божественный Чарли (2004) // «Экран и сцена», 2004.05.06][Пример из НКРЯ]

В примере (7) частица *как раз* устанавливает соответствие между фильмом «Огни рампы» и непредсказуемым фильмом, причем, согласно основному значению этой частицы, это соответствие рассматривается как нечто неожиданное, противоречащее какому-либо ожиданию (реальному или предполагаемому) слушающего. (Пайар 1997, 587).

Функция соответствия обычно способствует установлению связи с известными или актуализованными в тексте объектами и оказывается одним из признаков определенной референции:

(8) *Менее глобальные вещи,* ***как раз*** *те, которые Вы перечислили, у нас с сыном пройденный этап.* [Наши дети: Подростки (2004)] [омонимия снята]

Однако в некоторых случаях соответствие может устанавливаться не с объектом или лицом, а с фрагментом значения имени, в частности, признаком. И в этом случае *как раз* или *именно* может употребляться в контексте, типичном для неопределенной референции:

(9) *А в результате окажется, что мне нужен как раз новичок.*

Таким образом, частицы *как раз, именно, вот* могут использоваться в качестве вспомогательного, но не абсолютного средства выявления определенной референции. При заданном порядке исследования референциального статуса имени они должны включаться после выявления невозможности применения других показателей.

4.2. Те же дискурсивные слова, которые используются для влияния на восприятие денотативного статуса имени, могут относиться и к предикатам — глаголам и прилагательным и тоже определенным образом способствовать восприятию соответствующей информации о них. Рассмотрим одно из них.

Частица *вот* имеет несколько функций уже в роли усилительной частицы, помимо значений указательной частицы (или наречия), для этого слово первичных. Как мы отметили выше, существует еще одна возможность передавать при ее помощи информацию, важную для осуществления референции: указание на известность, активизированность слова в предыдущем (но, возможно, не очень близком) контексте. Сказанное может быть актуально и тогда, когда в сфере действия частицы *вот* оказываются глаголы.

Частица *вот* достаточно широко распространена в устной речи и иногда воспринимается как «слово-паразит» (хезитатор), т. е. средство создания паузы

в речи, необходимой для окончательного грамматического и лексического оформления высказывания:

(10) *Недавно я узнал, что машинисты и,* ***вот*** *эти, локомотивы, ну, то есть тепловозы, электровозы, ну* ***вот*** *эти,… на железной дороге…* [Евгений Гришковец. ОдноврЕмЕнно (2004)] [пример из НКРЯ]

Однако нередко *вот* используется с иной функцией. Так, имеются примеры, свидетельствующие о том, что частица *вот* способствует актуализации какого-то события, хранящегося в памяти собеседников:

(11) *Вернулся* ***вот*** *я из армии…*

Здесь частица обозначает, что предикат не несет новой информации, а является ссылкой на уже известное событие. Если бы в сообщаемой информации использовалось имя (*возвращение из армии*), то оно получило бы референциальный статус определенности.

Еще пример:

(12) *Что было — крутилось вкруг сетевых форумов и моих любимых космических опер. Писать я начал, но быстро понял, что ничего не выходит, кроме реверансов в сторону Кирилла Бенедиктова, Игоря Алимова, Маши Звездецкой, Олега Дивова «и так далее». Сказки я так и не написал, сослался на творческий крайзис. Зато сегодня* ***вот*** *обнаружил файл с загадочным названием «Ре». Дай, думаю, посмотрю, что это я. Посмотрел… :] 1.* [Запись LiveJournal (2004)] [пример из НКРЯ].

Поскольку интроспекция тоже является инструментом лингвиста, автор может привести пример из собственной активности в сетях. Выложив объявление о конференции, на следующий день он размещал пост, где та же конференция упоминалась. В этом случае показалось уместным (хотя и необязательным) упоминание конференции с частицей *вот*:

(13) *А вот где учат умению так ловко вывернуть сказанное* (комментатор сообщил, что в России диспансеризация фигуристов обязательна, в Америке — нет, вот и не выявили заболевание — и только!) *могу сообщить. Например, у нас в МГПУ. Правда, мы* ***вот*** *и конференцию по этико-правовым аспектам сейчас проводим, и вообще учим, чтобы поосторожнее https://www.facebook.com/profile.php?id=100001086466281&fref=pb&hc_location=friends_tab*

Очевидно, что если в первых двух случаях употребление *вот* вполне соответствует ее отмеченным значениям, последнее вхождение частицы ничем иным, кроме желания показать, что автор помнит об известности для читателей новости о проведении конференции, объяснить нельзя.

Еще более распространенным оказывается употребление *вот* в тех случаях, когда предикат сообщает об уже актуализованном в тексте действии:

(14) *И когда мы **вот** снимали, поняли…*

(Этот фрагмент реплики был зафиксирован в ходе рассказа о съемках роликах. К этому месту беседы то, что ролик снимался, уже было известно, несколько раз упомянуто).

Как отмечено в (Борисова 1998), в речи некоторых употребление *вот* для маркировки уже упоминавшегося действия может стать едва ли не обязательным. Однако в целом появление частицы *вот* зависит и от типа речи, и от идеолекта. Как и в других случаях, рассмотренных в данной статье, мы не можем говорить об обязательности использования частиц в описываемых случаях, речь идет только о возможности содействия тем или иным механизмам речевой деятельности, в том числе и связанным с референцией.

Аналогичную функцию могут выполнять и другие частицы:

(15) *Мы прекрасно понимаем, что это нелегко даётся. Это результат напряжённой последовательной работы, Вашей и Ваших коллег. Но мы рады Вашим успехам и готовы со своей стороны всячески содействовать стабильному развитию Йемена, стабильному развитию обстановки в регионе. Надеемся, что сотрудничество между нашими странами **как раз** и будет служить этой цели. С. В. Лавров. Встреча министров иностранных дел прикаспийских государств. Выступление на пресс-конференции по итогам встречи // «Дипломатический вестник», 2004*[пример из НКРЯ]

В этом фрагменте речи *как раз* показывает, что о служении «этой цели», т. е. стабильному развитию обстановки, уже говорилось

5. Способы указания на неопределенного референта

Если в предыдущих параграфах мы имели дело со сложностями при указании на определенную референцию, то сейчас коснемся проблем, возникающих тогда, когда автор подразумевает неопределенность объекта.

В безартиклевых языках имя без указательных частиц и других средств определенной референции (например, если это не имя собственное) далеко не всегда может пониматься как носитель неопределенной референции. Если его актуализация не обеспечена предшествующим контекстом или другими средствами (в том числе, и рассмотренными выше), слушающий может понять имя как указание на принадлежность к классу. В принципе, это классический способ неопределенной референции — указание на неизвестный объект через принадлежность его к классу:

(16) *А меня милиционер через дорогу перевел.*

Однако по законам текстообразования далеко не любое указание на класс способно ввести неопределенный, ранее не упоминавшийся объект. В примере (15) милиционер попадает во фрейм «переход дороги», и его упоминание достаточно для неопределенной референции. Однако если такого попадания во фрейм нет, указание на принадлежность к классу возможно только при наличии дополнительных маркеров.

(17) *И откуда у тебя сведения? — Да мне* ***там*** *студент подсказал* (из беседы преподавателей)

В этом случае становится необходимо задать эту неопределенность. Обычно это делается при помощи местоимения *один* (Фёдорова 1999). Наряду с ним или даже вместе с ним может использоваться дискурсивное слово отвлечения внимания *там*:

(18) *Еще гад* ***там один*** *«пошутил», добавил мне переживаний.* [Алексей Буданов. Паттайа для белого человека // «Пятое измерение», 2003] [пример из НКРЯ].

Частица *там* используется для указания на неважность сообщения и тогда, когда имя имеет определенный статус:

(19) *Да,* (машет рукой) *он* ***там*** *с женой не поладил.*

Однако в случае отсутствия сведений об определенности именно указание на неважность происхождения объекта, его свойств, помогает представить этот объект как неопределенный:

(20) • ***пчёлка Майя*** *, 2004. 07. 1212: 26. Подскажите, плз. приблизительно, когда мальчиков начинают интересовать девочки?? Ну* ***там*** *звонки, записочки на уроках..: -)* ***Fantom*** *, 2004. 07. 1213: 05. Мне вот тоже хочется знать: -).* [Наши дети: Подростки (2004)] [пример из НКРЯ]

(21) *А вот совсем рядом, в Африке, я сам видел, местные царьки-президенты продают транснациональным корпорациям недра и себя самих в качестве вменённого налога, безо всякой* ***там*** *демократии и без группы ловких юношей, которые её олицетворяли бы. И без шестидесятников, которые бы этих юношей защищали. Проще. Прямее.* [Сергей Доренко. Левые силы — перезагрузка (2003) // «Завтра», 2003.08.13] [пример из НКРЯ]].

Заметим, что добавление частицы *там* практически исключает одну из разновидностей неопределенной референции, а именно интродуктивную, когда употребление в тексте первый раз превращает объект в известный, т. е. этот объект вводится в число определенных. *Там* показывает, что дальнейшего

внимания этому объекту уделяться не должно. Если и находятся исключения, то они могут объясняться сменой тактики речевой деятельности:

(22) *Мне **там** одна дама посоветовала на рынке не брать. Да ты, может, ее знаешь? Соседка с первого этажа. Ну вот, мне она….*

6. Референция и другие механизмы понимания

Из сказанного можно сделать вывод, что нередко дискурсивные слова участвуют в процессе передачи сведений о денотативном (референциальном) статусе имен в высказываниях. При этом они не только сохраняют лексическое единство с теми же словами в значениях усиления, подчеркивания и т. п. Можно сказать, что и в рассмотренных нами случаях взаимодействия со средствами референции дискурсивные слова сохраняют ту же функцию, которую они имеют в других употреблениях, неважных для референции.

Следовательно, основания для взаимодействия связаны не с выработкой особых значений или назначений дискурсивных слов, а с речевыми процессами, в которых аспекты, использующие маркировку частицами: известность, актуализованность, ожидаемость и т. п. — тесно переплетаются с процессами установления референции.

Дискурсивные слова маркируют различные процессы понимания. В частности, они используются для усиления внимания к объекту или, напротив, ослаблению, При этом усиление внимания предполагает облегчение поиска денотата в отражаемом мире.

Рассмотрение их употребления позволяют увидеть то общее, что есть у референции и других процессов.

Это поиск объекта в ситуации резких сдвигов (начальное обращение, приведение примера), указание на отсутствие «предмета прикрепления» при неопределенном употреблении, указание на совпадение с известным (обычно уже упоминавшимся в данном тексте) объектом или действием.

Просматривается и связь с коммуникативной организацией высказывания, что уже давно было установлено для дискурсивных слов, с одной стороны, и для определения денотативного статуса имен, с другой.

Таким образом, обращение к процессам понимания, взятым комплексно, с учетом различных сторон этого процесса, может облегчить анализ отдельных аспектов высказывания. И учет особенностей дискурсивных слов может помочь в установлении референциального статуса имен и некоторых, близких к нему, характеристик предикатов.

Литература

1. *Борисова Е. Г.* (1982) Семантический анализ усилительных частиц русского языка. — Автореф… канд. филол. н. М. 1982.
2. *Борисова Е. Г.* (1998) Управление вниманием говорящего при помощи частиц // Труды Международного семинара «Диалог-98» по компьютерной лингвистике и ее приложениям, Казань, 1998 г. — сс. 84–88.
3. *Борисова Е. Г., Овчинникова Т. Е.* (2005) Пространства усиления (пространственная метафора и возникновение усилительных частиц)// Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2005»/ под ред И. М. Кобозевой, А. С. Нариньяни, В. П. Селегея. — М.: «Наука»
4. *Падучева Е. В.* (1985) Высказывание и его соотнесенность с действительностью (референциальные аспекты семантики местоимений). — М: «Наука».
5. *Пайар Д.* Формальное описание дискурсивных слов: как раз и именно / Дени Пайар // Formale Slavistik. — Leipzig: University of Leipzig, 1997. — С. 586–593.
6. *Федорова О. В.* Неопределенное местоимение ОДИН в русском языке как показатель интродуктивной референции имени // Вестн. Моск. ун-та. Сер. 9, Филология. — М., 1999. — N 2. — С. 98–112.

References

1. *Borisova E. G.* (1982) Semantic Analysis of Modal Particles in Russian PhD Thesis. Moscow. [Semanticheskij analiz usilitel'nyx chastits russkogo jazyka.]
2. *Borisova E. G.* (1998) Governing of the Attention of the Speaker with the Help of Modal Particles [Upravlenije vnimaniem govoryaschego pri pomoschi usilitel'nyx chastits / Proceedings of the International Conference "Dialog 98" Kazan.
3. *Borisova E. G., Ovchinnikova T. E.* (2005) The Spaces of Emphasizing (the space Metaphor and Rise of Modal Particles], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2005" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2005"], Bekasovo,
4. *Fedorova O. V.* (1999) Undefinite Russian pronoun ODIN as the marker of the introductive function of the noun reference [Neopredeljonnoe mestoimenije ODIN kak marker introduktivnoj funktsii] Vestnik Moskowskogo Universiteta. Ser 9 Filologija — M., 1999. — N 2. — P. 98–112.
5. *Paducheva E. V.* (1985) Utterance and its Reference to Reality [Vyskazyvanie i ego sootnesënnost` s dejstvitel`nostju].
6. *Paillard D.* (1997) Formal Description of Discourse words kak raz and imenno // Formale Slavistik. — Leipzig: University of Leipzig, 1997. — P. 586–593

ONTOLOGY AND INTEGRATION OF FORMAL AND LEXICAL SEMANTICS

Borschev V. B. (borschev@linguist.umass.edu),
Partee B. H. (partee@linguist.umass.edu)

University of Massachusetts, Amherst, USA

Formal and lexical semantics can be integrated if they speak the same language. We claim that a substantial part of lexical semantics can be incorporated into formal semantics without adding to the latter any new mechanisms. This talk continues the authors' work on the ontology and the semantics of measure constructions in Russian. The work concerns expressions like *dva stakana moloka*, *polkorziny gribov, tri meshka muki* (*two glasses of milk, half a basket of mushrooms, three bags of flour*), etc., describing various kinds of *containers*, or corresponding measures based on them, and their contents—*portions of substances*. In our previous works, describing ontological information, including *sorts* of things and the words and expressions that designate sorts, we did not include those sorts in our formal semantic analyses. We do that in the present work, declaring *sorts* as *types* and thereby significantly expanding Montague's system of types. On the one hand this gives us the means for specifying various aspects of the ontology, and on the other hand it lets us more fully specify the semantics of the constructions under consideration. The substantive goals of this research are, in part, to be able to describe and explain co-occurrence constraints and ideally to be able to formally distinguish well-formed from ill-formed expressions in this domain.

Keywords: formal and lexical semantics, ontology, types and sorts, genitive of measure

1. Introduction

Our Main Thesis

A substantial part of lexical semantics can be incorporated into formal semantics without requiring the addition of any new mechanisms to formal semantic theory.

Background

This talk continues the authors' work on the ontology and the semantics of measure constructions in Russian[1]. The work concerns expressions like *dva stakana moloka* 'two glasses of milk', *polkorziny gribov* 'half a basket of mushrooms', *tri*

[1] See Borschev, Partee 1999, 2001a, 2001b, 2011; Partee, Borschev 2010, 2012a, 2012b; and also Borschev 2014a, 2014b. For the first author, the background of this work is also connected with works of his late wife L. V. Knorina on the semantics of the genitive construction. Some results were reported in Borschev, Knorina 1990.

meshka muki 'three bags of flour', etc., describing various kinds of *containers*, or corresponding measures based on them, which we will call *container-measures*, and their contents—*portions of substances*.

In our previous works, describing ontological information, including *sorts* of things and the words and expressions that designate sorts, we did not include those sorts in our formal semantic analyses. We do that in the present work, declaring *sorts* as *types* and thereby significantly expanding Montague's system of types. On the one hand this gives us the means for specifying various aspects of the ontology, and on the other hand it lets us more fully specify the semantics of the constructions under consideration.

The substantive goals of this research are, in part, to be able to describe and explain co-occurrence constraints and ideally to be able to formally distinguish well-formed from ill-formed expressions in this domain.

Examples. A fragment of ontology for the expression of measure

(1) *On vypil dva stakana moloka.*
He drank two glass-GEN.SG milk-GEN.SG
He drank two glasses of milk.

(2) *Voz'mite poltora stakana muki.*
Take one-and-a-half glass-GEN.SG flour-GEN.SG
Take one and a half glasses of flour.

(3) *On prines polkorziny gribov.*
He brought half-basket mushroom-GEN.PL
He brought half a basket of mushrooms.

(4) *dva puchka rediski*
two bunch-GEN.SG radish-GEN.SG
two bunches of radishes

The examples above contain the *genitive measure construction.* In the first three, the measure is constituted by *containers*—in this case glasses and baskets; these are *container-measures*. Jars, bags, boxes, etc., can also be used as measures. They can be *filled*—completely or to some degree—with various *substances*—milk, water, flour, mushrooms, etc., and so can serve as a measure of quantity of such substances. The contrasting example (4) is a genitive measure construction but does not use a container-measure.

Substances are of various sorts—*liquids, granular substances,* and others. In our examples we are concerned with the measuring of *portions of substances*.

We note that formal semanticists have discussed ontological and semantic commonalities in the description of plural entities and mass stuff in natural language; see Parsons 1970, Link 1983, Landman 2004. The normal quantity measure for finite pluralities of entities is their cardinality—the number of elements in the corresponding

set, which is a whole number (*five boys*)[2], and a normal measure for portions of substances is their volume[3], measured in terms of certain standard portions (*two liters of milk* or *one and a half cups of flour*), and in this case we find fractional as well as whole numbers.

Having taken the quantity of a certain portion as a unit, that is, as a unit of measure, we can measure portions of substances in those units, for instance in liters or in the volume of particular glasses, baskets, and the like, determining how many liters (glasses, baskets) or parts thereof are contained in a given portion of substance. Therefore with each unit of measure there is a correlated *function*, defined on portions of substances. For example, corresponding to the liter unit we can assign the function LITER: if ***m*** is a portion of milk, then LITER(***m***) is a number giving the volume of that portion in liters (cf. Landman 2004).

The preceding text could be called a 'dotted-line outline' of a fragment of ontology for expressions of container-measure. Ontology, within philosophy, is a branch of metaphysics that studies what there is and the nature of the basic categories of the things that make up the world. The task of natural language metaphysics (Bach 1986a) is to understand what presuppositions a language makes about how the world is structured, and natural language ontology is a part of that task.

Ontology studies the various kinds of existents, and usually includes some classification of their sorts and types. *Glasses (cups), baskets, bags, containers, water, milk, flour, liquids, granular matter, substances*—these are examples of *sorts* of entities which we have considered in our previous works.

The semantics of expressions of natural languages rests on ontology. The examples mentioned so far are semantically well-formed, because they rest on an ontologically correct picture of the world.

Thus in example (1), milk is a liquid, a substance, it can fill a glass, and all glasses are containers. And since milk is a liquid, a portion of substance constituting two glasses of milk is something that can be drunk.

In an analogical manner the well-formedness of expressions (2) and (3) is based on ontology. The bunch of radishes in example (4) is a natural 'portion' of radishes; one can measure radishes in bunches and count the bunches.

But examples (5–7) below are ontologically ill-formed, or at least doubtful[4].

(5) *# On vypil dva stakana muki.*
He drank two glass-GEN.SG flour-GEN.SG
He drank two glasses of flour.

---

[2] Some count nouns denote objects that are divisible: *a half of pie, one and a half of an apple*. Such expressions do not denote simple 'pluralities' or 'sets', and we do not discuss them here.

[3] Portions of substances can also be measured by weight or mass, but we restrict our attention here to volume.

[4] We use the symbol # to mean 'anomalous', without specifying whether the anomaly should count as syntactic, semantic, or pragmatic.

(6) *# dva puchka moloka*
two bunch-GEN.SG milk-GEN.SG
two bunches of milk

(7) *??On uronil s podnosa poltora stakana moloka.*
He dropped from tray one-and-a-half glass-GEN.SG milk-GEN.SG
He dropped from the tray one and a half glasses of milk.

Dva stakana muki 'two glasses of flour' is well-formed: flour, like other particulate matter, can be measured in glasses and one can count the corresponding portions. So two glasses of flour is a quantity of flour. But flour cannot be drunk[5]; one can drink only liquids.

And example (6) is ill-formed because portions of liquid aren't the kind of thing that can occur in bunches.

Things are somewhat more complex with example (7). *One and a half glasses of milk* is a well-formed expression denoting a portion of milk measuring one and a half glasses. But the verb *ronjat'* ('drop') does not apply to portions of substance, but to objects. For a liquid an appropriate verb would be *prolit'* ('spill'). And on a tray, we carry objects, not (directly) portions of liquid. And while portions of matter can be measured in fractional container-measures—one and a half glasses, half a basket, etc., normal objects are counted with whole numbers. Half of a portion of milk is also a portion of milk, but half a chair is not a chair and half a glass is not a glass. Hence one cannot drop one and a half glasses of milk from a tray.

Our task here is to formally describe a fragment of the ontology of natural language on which the semantics of measure expressions depends. We aim to do that by giving a semantics that assigns suitable meanings for semantically well-formed expressions and accounts for the anomaly of semantically ill-formed expressions.

Our larger goal is to show that this can be done with the tools of formal semantics if we include in formal semantics at least a certain part of lexical semantics.

2. Formal Semantics

Formal semantics of natural language is historically associated with the name of R. Montague. Montague showed that the syntax and semantics of natural language can be described using the tools developed by logicians for the formal description of their formal languages. These methods give a model-theoretic semantic interpretation of syntactic structures, obeying the principle of compositionality. The tools for

[5] We return at the very end to the question of whether the prohibition on drinking flour is really a matter of semantic ill-formedness; one can argue against that idea from the well-formedness of expressions like "one cannot drink flour", which are completely well-formed and understandable and which have "drink flour" as a subpart. But for now we treat (4) as semantically ill-formed.

such formal description have been greatly extended in the last forty years by the cooperative efforts of linguists, logicians, and philosophers of language.

Over the last forty-plus years formal semantics has become (especially in the West) the mainstream approach to semantic research.

But especially in the beginnings, formal semantics by no means described the whole semantics of natural language. Montague did not try to describe lexical semantics, considering that a more empirical task. Montague's semantics can be reasonably characterized as *the semantics of syntax* (Paducheva's term).

Formal semanticists are always thinking about compositionality, how the meaning of a sentence (or any other complex expression) is built up from the meanings of its parts. And on the one hand, this requires having some ideas about the meanings of the smallest parts—words and morphemes—because they form the starting point for semantic composition. So formal semantics needs some kind of lexical semantics to start from. The bare minimum is to make some assumptions about the nature of lexical meanings and not make any specific claims about any particular lexical meanings—that was Montague's strategy, since he had neither the interest nor the competence to address empirical matters of lexical semantics. He limited himself to trying to figure out the "semantic type" of various classes of lexical items, and the actual semantics for certain key 'logical words'.

Montague's framework uses two basic types: **e** and **t**, and every model includes two basic domains, $\mathbf{D}_e$ and $\mathbf{D}_t$, the set of all 'entities' of the given model and the set of truth values (normally1 and 0.) 'Entity' here is considered in the broadest possible way, including ordinary objects as well as numbers, colors, wars—anything a language has names for. (Semanticists have sometimes added additional basic types, for instance for events or situations, for moments or intervals of time, for degrees (used in the semantics of comparatives and other degree modification), for numbers.)

Starting from the basic types, a hierarchy (tower) of functional types is constructed: <**e**,**t**>, <**e**,**e**>, <<**e**,**t**>, **t**> etc. The type <**a**,**b**> corresponds to the domain $\mathbf{D}_{<a,b>}$, the set of all functions *f* from domain $\mathbf{D}_a$ to domain $\mathbf{D}_b$.

With this hierarchy of types, Montague has a framework for an important part of the ontology of natural language, namely providing semantic values for expressions of all sorts of syntactic categories, including everything from nouns, verbs and adjectives to adverbs, declarative and interrogative sentences, embedded propositions, etc. And since the most basic way that expressions combine semantically is by function-argument application, this simple type structure characterizes which expressions can combine with which others. Within this relatively simple ontology he is thus able to capture the "semantics of syntax."

Since its beginnings around 1970, there has been a great deal of work in formal semantics, including work that brings formal semantics and lexical semantics together. We have already mentioned some of the work of Parsons, Link, and Landman. We also note the work of E. Bach (Bach 1986a, 1986b) on natural language metaphysics. There are many other works which have extended formal semantics by including more lexical semantics and making use of ontological specifications, including Dowty 1979, Kamp and Partee 1995, Pustejovsky 1995.

3. What's New in This Paper?

Building on our previous work, we refine our earlier semantics for measure constructions in Russian, adding ontological information. Technically this is accomplished by unifying the notions of *type* and *sort*: the sorts to which words and phrases belong become types and are added to the hierarchy of types. This radically increases the collection of types, and a significant part of lexical semantics immediately becomes part of formal semantics.

So for the measure expressions which are considered in this work, we introduce new basic types for different kinds of containers and portions of substance: **glass**, **basket**, …, **container**, and also **milk**, **water**, **flour**, … , **liquid**, **granul_subst**, **pourable_subst**.

For these types we introduce the corresponding domains $\mathbf{D}_{glass}$, $\mathbf{D}_{basket}$, $\mathbf{D}_{container}$, and likewise $\mathbf{D}_{milk}$, $\mathbf{D}_{water}$, $\mathbf{D}_{flour}$, …, $\mathbf{D}_{liquid}$, $\mathbf{D}_{granul\_subst}$, $\mathbf{D}_{pourable\_subst}$, etc.

The domain $\mathbf{D}_{glass}$ is the set of all glasses, $\mathbf{D}_{basket}$ is the set of all baskets, and $\mathbf{D}_{container}$ is the set of all containers, including all glasses, baskets, etc. So every glass is included in the domain corresponding to the type **glass**, every basket similarly corresponds to the type **basket**, and both glasses and baskets also correspond to the type **container**; $\mathbf{D}_{container} = \mathbf{D}_{glass} \cup \mathbf{D}_{basket} \cup \mathbf{D}_{bag} \cup \ldots$.

Formally all the domains we have just identified are subsets of the domain $\mathbf{D}_e$ and are picked out by characteristic functions from $\mathbf{D}_e$ to $\mathbf{D}_t$, the same kinds of functions that correspond to one-place predicates of entities. Thus the domain $\mathbf{D}_{glass}$ is formally defined by the predicate ***glass*** of type <e,t>, whose characteristic function from $\mathbf{D}_e$ to $\mathbf{D}_t$ yields the value **1** (true) for all glasses and **0** (false) for all other entities in $\mathbf{D}_e$.

In some sense the opposite may be true: in our linguistic consciousness the domain $\mathbf{D}_e$ is probably a generalization that is derived from its more familiar subdomains.

The situation with substances is analogous. So the domain $\mathbf{D}_{milk}$, corresponding to the type **milk**, consists of all portions of milk; the corresponding characteristic function is ***milk***: $\mathbf{D}_e \rightarrow \mathbf{D}_t$, yielding the value **1** for all portions of milk. And to the type **liquid** there corresponds the domain $\mathbf{D}_{liquid}$, consisting of all portions of liquid, and of course $\mathbf{D}_{milk} \subset \mathbf{D}_{liquid} \subset \mathbf{D}_{pourable\_subst}$. In exactly the same way we have $\mathbf{D}_{flour} \subset \mathbf{D}_{granul\_subst} \subset \mathbf{D}_{pourable\_subst}$, since both liquids and granular substances are pourable substances. The domains $\mathbf{D}_{liquid}$ and $\mathbf{D}_{granul\_subst}$ do not intersect. These are all parts of the naive ontology that makes up part of the naive metaphysics of every language user.

Following Landman 2004 we also introduce type **r** as the type of real numbers, and the corresponding domain $\mathbf{D}_r$.

From the "new" basic types, as well as the basic types **e** and **t**, we build the hierarchy of types. And we will consider functions whose arguments and values fall within these domains.

Now we can say that the already mentioned function constant LITER belongs to the type <**pourable_subst, r**>, denoting a function in the domain $\mathbf{D}_{<pourable\_subst, r>}$, i.e., $\mathbf{D}_{pourable\_subst} \rightarrow \mathbf{D}_r$. Given the natural partial order among sorts of substances, the function LITER is defined not only on for arguments in the domain $\mathbf{D}_{pourable\_subst}$, but also on the domains $\mathbf{D}_{milk}$ и $\mathbf{D}_{liquid}$, and also on the domain $\mathbf{D}_{granul\_subst}$, but it is not defined on bunches of radishes. In general when we assign a type <**a,b**> to a lexical

item, the function that is the semantic value of that lexical item will be defined only for arguments in domains that have a non-empty intersection with $\mathbf{D}_a$.

Two Observations

1. In introducing new types, relations among them, and certain functions defined on them, we obtain the means for describing a fragment of the ontology of natural language.
2. Our modifications to the system create *multi-sorted models*, that is, models in which there are many basic domains. Multi-sorted systems are introduced in order to delimit the domains of definedness of functions[6], as for instance in our example function LITER.

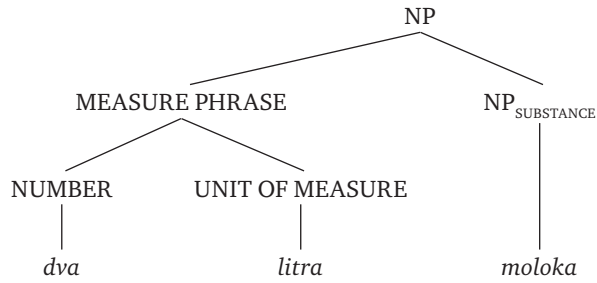
4. Examples

In Partee, Borschev 2012b we described the semantics of the expressions *dva litra moloka* 'two liters of milk' and *dva stakana moloka* 'two glasses of milk'. Here we show how that description is modified when we make use of the new types introduced here.

Dva litra moloka 'two liters of milk'

First we present the syntactic structure of the expression in tree form.

(8)



Let us begin with *litra*. We introduced above a constant LITER of type <**e**,**r**>, more precisely <**pourable_subst**, **r**>, denoting a function that maps a portion of any 'pourable substance' onto a number that gives its volume in liters. For the use of *litr* in the genitive of measure construction, as it occurs in (8), we make use of a derived function constant LITER_2 defined in terms of LITER. LITER_2 takes a number *n* as argument and returns a predicate modifier, a function that can apply to the semantic value

[6] An alternative way to delimit the domains of functions is to introduce presuppositional conditions in the definition of the particular function, so that the function is restricted to applying to only a subpart of the domain specified by its type. Such stipulated restriction on the domain can be used for restrictions other than sortal ones. We leave comparison of these approaches for future research.

of the NP *moloka* to return a predicate true of anything which is a portion of milk and has a volume of *n* liters, letting us express *so-and-so many liters of such-and-such substance*.

Definition: $\text{LITER}_2 = \lambda n\ [\lambda P\ [\lambda x\ [(\text{LITER}(x) = n)\ \&\ P(x)]]]$.

Here the first argument, n, of LITER_2 is a variable of type **r** over numbers; the second argument, P, is of type <**e,t**>. Because the original function LITER is defined only for entities of type **e** that are furthermore of type **pourable_subst**, and since according to the formula in the definition of LITER_2, P must apply to x, which is also an argument of LITER in the same formula, the only admissible values for P will be properties that can hold of an entity x which is of the type **pourable_subst**.

In terms of Montague's basic type structure (assuming that numbers are a subset of entities), the type of the variable n above is **e**, the type of P is <**e,t**>, and the type of x is **e**. The type of the whole formula to the right of the = sign, and hence of LITER_2, is <**e**, <<**e,t**>, <**e,t**>>>: it maps an entity (a number) onto a function from properties to properties.

Using our enriched type system that includes sortal information, we can specify the types of n, P, and x more narrowly. We have already noted that n is of type **r**. As a result of the constraints imposed by LITER, we can determine that any admissible value for x in the formula must be of the type **pourable_subst** (or some subtype thereof; what LITER tells us is that any value of x must be of a type that is compatible with the type **pourable_subst**). And since, as noted above, P must apply to that x, any admissible value for P must be of type **<pourable_subst, t>**. We thus derive that the more fine-grained type of LITER_2, is <**r**, <<**pourable_subst, t**>, <**pourable_subst, t**>>>: it maps a number onto a function from properties of pourable substances to properties of pourable substances, letting us express *so-and-so many liters of such-and-such substance*.

Below we spell out the semantic derivation for the expression *dva litra moloka*.

(i) *litr: litr*$_2$: Type <**r, <<pourable_subst, t>, <pourable_subst, t>>>**
Meaning: $\text{LITER}_2 = \lambda n\ [\lambda P\ [\lambda x\ [(\text{LITER}(x) = n)\ \&\ P(x)]]]$
(ii) *dva*: Type **r**. Meaning: 2.
(iii) *dva litra*: Type **<<pourable_subst, t>,<pourable_subst, t>>**
Meaning: $\lambda P\ [\lambda x\ [(\text{LITER}(x) = 2)\ \&\ P(x)]]$
(iv) *moloka*: Type <**milk**,**t**>. Meaning: ***milk***
(v) *dva litra moloka*: Type <**milk**,**t**>.
Meaning: $\lambda P\ [\lambda x\ [(\text{LITER}(x) = 2)\ \&\ P(x)]]$ (***milk***) $= \lambda x\ [(\text{LITER}(x) = 2)\ \&$ ***milk*** $(x)]$

Note that according to line (iii), any admissible argument of *dva litra* must be of the type **<pourable_subst, t>**. Since **milk** is a subtype of **pourable_subst,** *moloka* is an admissible argument for *dva litra*.

How do we determine that the sort, or type, of the result has the more narrow specification <**milk**,**t**> rather than the more inclusive sort **<pourable_subst, t>**? That follows from the fact that the interpretation of the result includes the subformula

milk (x); therefore any admissible value of x in the interpretation of *dva litra moloka* must satisfy the more restrictive constraint that it be of sort **milk** and not merely the constraint imposed by the subformula "LITER(x) = 2" that it be of sort **<pourable_subst, t>**.

And note, as we will illustrate below, that if the sortal restrictions imposed by the two subformulas were not compatible, the whole expression would be semantically ill-formed.

***Dva stakana moloka* 'two glasses of milk'**

In Partee, Borschev 2012b we examined several variants of the semantics of this expression, relating them by some container-specific meaning-shifting rules. Here, because the shifting rules are not our center of interest, we will apply our sortal approach to just one of them—what we called the "Ad Hoc Measure" reading, in which a concrete glass of arbitrary size is used to provide a unit of measure (*stakan*$_{AHM1}$ in the terminology of the cited work).

In that interpretation, the word *stakan* has undergone a lexical shift—it denotes not some concrete glass ***c***, but a unit of measure of substances, corresponding to the volume of a portion of substance that would fill this concrete glass ***c*** and analogous to other units like *liter* and *pint*.

Just as with LITER, and with any term for a unit of measure for measuring volumes of substances, we will have both a basic function denoted by STAKAN$_{AHM}$ and a derivative term STAKAN$_{AHM2}$ that will be used in the genitive of measure construction.

STAKAN$_{AHM}$: Type <**pourable_subst**,**r**>. Meaning: the denotation of STAKAN$_{AHM}$ is a function from $\mathbf{D}_{pourable\_subst}$ to $\mathbf{D}_{r}$ corresponding to some concrete glass ***c***, such that if ***m*** is a portion of substance, then STAKAN$_{AHM}$(***m***) is the volume of ***m*** measured in terms of glass ***c***.

The derived STAKAN$_{AHM2}$ used in the genitive construction has an argument structure like that of LITER$_2$, letting us express *so-and-so many glasses of such-and-such substance:*

$$\text{STAKAN}_{AHM2} = \lambda n\,[\lambda P\,[\lambda x\,[(\text{STAKAN}_{AHM}(x) = n)\ \&\ P(x)]]].$$

And the semantic derivation for an expression like *dva stakana moloka* for an arbitrary glass is completely analogous to the semantic derivation for the expression *dva litra moloka*.

(i) *stakan: stakan*$_{AHM2}$: Type <**r, <<pourable_subst, t>, <pourable_subst, t>>>**
Meaning: $\text{STAKAN}_{AHM2} = \lambda n\,[\lambda P\,[\lambda x\,[(\text{CTAKAH}_{AHM}(x) = n)\ \&\ P(x)]]]$.

(ii) *dva*: Type **r**. Meaning: 2.

(iii) *dva stakana*: Type **<<pourable_subst, t> ,<pourable_subst, t>>**
Meaning: $\lambda P\,[\lambda x\,[(\text{STAKAN}_{AHM}(x) = 2)\ \&\ P(x)]]$.

(iv) *moloka*: Type <**milk**,**t**>. Meaning: ***milk***

(v) *dva stakana moloka*: Type <**milk**,**t**>.
Meaning: $\lambda P\,[\lambda x\,[(\text{STAKAN}_{AHM}(x) = 2)\ \&\ P(x)]](\textbf{\textit{milk}}) =$
$\lambda x\,[(\text{STAKAN}_{AHM}(x) = 2)\ \&\ \textbf{\textit{milk}}(x)]$.

***Dva stakana* (*kakoj-to*) *gadosti* 'two glasses of (some) filth'**

Gadost' is an evaluative word, and like English *filth*, or *nasty stuff*, it can be applied to things and stuff and happenings of the most varied sorts. Someone can speak *gadost'* to us, or do *gadost'* to us; a movie or a situation can be called *gadost'*, etc. In the genitive of measure construction in example (9), the 'filth' must be understood as some substance which can be measured by the glassful and can be drunk, hence some liquid; that follows from the sortal requirements of the other parts of the construction and the semantics of the construction.

(9) *On vypil dva stakana (kakoj-to) gadosti.*
He drank two glass-GEN.SG (some-kind-of-GEN.SG) filth-GEN.SG
He drank two glasses of (some sort of) filth.

We start with a minimal Montagovian representation of the interpretation of *gadost'* as a predicate **filth'**, and assign to it the inclusive predicate type <e,t>, since evaluative predicates do not generally have sortal restrictions. All we need to know for this example is that the type for *gadost'* has a non-empty intersection with the types **<pourable_subst, t>** and **<liquid, t>**.

Then the semantic derivation for the expression *dva stakana* (*kakoj-to*) *gadosti* will be similar to that of the expression *dva stakana moloka*, differing only in when and how the semantic type of the result is determined. The first three lines, (i)-(iii), which just concern *dva stakana*, will be identical, so we just provide steps (iv) and (v) below.

(iv) *gadosti*: Type **<e, t>**. Meaning: ***filth***'
(v) *dva stakana (kakoj-to) gadosti*: Type **<pourable_subst, t>**.
Meaning: $\lambda P\ [\lambda x\ [(\text{STAKAN}_{AHM}(x) = 2)\ \&\ P(x)]]\ (\textbf{\textit{filth'}}) =$
$\lambda x\ [(\text{STAKAN}_{AHM}(x) = 2)\ \&\ \textbf{\textit{filth'}}(x)]$

The type of the whole expression results from the restrictions imposed on the type of admissible values of x by the type of CTAKAH_{AHM}. Since ***filth'*** imposes no sortal restrictions of its own, the restrictions imposed by CTAKAH_{AHM} determine the final result.

5. Anomalous Examples

Let us once more contrast the "well-formed" and "ill-formed" examples from the Introduction. The expression *dva stakana moloka* from example (1) *On vypil dva stakana moloka* 'He drank two glasses of milk' can describe some portion of milk and is of type <**milk**, **t**>. The verb *vypit'* (*drink*) is defined for direct objects of the type **liquid**. The whole expression will require that what was drunk is both of type **liquid** and of type **milk** (because in the derivation of the meaning some e-type variable x will occur both as an argument of *vypit'* 'drink' and as an argument of the predicate ***milk***. And since **milk** is a subtype of **liquid**, that is consistent.

The expression *dva stakana muki* 'two glasses of flour' in example (5) is of type <**flour**, **t**>, the intersection of types <**flour**, **t**> and <**pourable_subst**, **t**>. But

<**flour**, **t**>, which is a subtype of <**granul_subst,t**>, is disjoint from the type <**liquid, t**>, and hence inadmissible as an argument of *vypit'* 'drink'. As a result, example (5) is semantically anomalous.

Example (6) (*Dva pučka moloka* 'two bunches of milk') is anomalous because *pučok* 'bunch' is not a unit of measure for portions of liquid, of type <**liquid,t**>.

Analogously, example (7) (??*On uronil s podnosa poltora stakana moloka* 'He dropped from the tray one and a half glasses of milk') is anomalous, or at least doubtful, since *uronit'* is restricted (in its literal uses) to solid objects and does not apply to portions of liquid—arguments of type <**liquid,t**>. And while *stakan* in its most basic use is a solid object, and some uses of *stakan moloka* do refer to the glass together with its contents (see Partee, Borschev 2012b), on those uses glasses can only be counted with whole numbers. When *stakan* has a measure interpretation as in (7), it can be measured in fractional numbers, but then the sort of the whole expression is <**pour_subst, t**>, not <**solid_entity**, **t**>. So it is impossible or nearly so to impose a consistent typing on the whole sentence in example (7).

One interesting complication, mentioned briefly in a footnote above, is that the restrictions we have explored hold for normal non-modal affirmative sentences. But in modal, interrogative, negative, and fictional contexts, these constraints do not always hold. Sentences like (10) and the English example (11) (from Thomason 1972) are fully acceptable.

(10) *Vrjad li on mog vypit' dva stakana muki.*
Hardly he could drink-PF two glass-GEN.SG flour-GEN.SG
It's doubtful that he could drink two glasses of flour.

(11) *It is not true that The Painted Desert is reluctant.*

These sentences contain subparts which we have analyzed as sortally incorrect. The conclusion should probably be that we want to use sortal information to explain the anomaly of the anomalous examples, but we do not want the semantic derivations to be impossible; we need the grammar to be able to generate the anomalous examples and the semantics to interpret them, so that they are available to be embedded under modals, negation, etc. Such examples, as Thomason argued, put some constraints on the nature of the explanation of sortal incorrectness. Making sort theory an extension of type theory, as we have done here, may in the end not be the best way to incorporate ontological information into the semantics.

* * *

We note in closing that even for the small fragment of ontology that we have considered here, the work is by far not complete. It is obvious that there are many and varied problems that arise. Words and expressions can belong to several types at once; one needs a mechanism for describing regular metonymy, metaphor, and other kinds of semantic shifts; the distinctions between words that belong to "ordinary"

ontological sorts and words like the evaluative *gadost'* 'filth' need to be studied; and there are many other problems.

These and other problems are beginning to receive greater discussion in works aimed at the integration of lexical and formal semantics. Interesting work of this sort is also going on in the context of advances in computational semantics. One can hope that with solutions to these problems and further such advances, formal semantics can progress from being the semantics of syntax to being a more complete semantics of natural language.

References

1. *Apresjan Ju. D.* (1999), Semantic motivation of nonsemantic properties of lexemes [Semanticheskaja motivacija nesemanticheskix svojstv leksem], In Die Grammatischen Korrelationen, Institut für Slavistik, Graz, pp. 96–116.
2. *Bach E.* (1986a) Natural language metaphysics, In Logic, Methodology, and Philosophy of Science VII, North-Holland, Amsterdam, pp. 573–595.
3. *Bach E.* (1986b) The algebra of events. Linguistics and Philosophy 9:5–16.
4. *Borschev V. B., Knorina L. V.* (1990), Types of entities and their perception in language [Tipy realij i ix jazykovoe vosprijatie], In Language of Logic and Logic of Language, Akademija Nauk SSSR, Naučnyj Sovet po Kompleksnoj Probleme "Kibernetika", Moscow, pp. 106–134.
5. *Borschev V. B.* (1996), Natural language as naïve mathematics [Estestvennyi jazyk—naivnaja matematika dlja opisanija naivnoj kartiny mira], in Moscow Linguistic Almanac 1, Jazyki russkoj kul'tury, Moscow, pp. 203–225.
6. *Borschev V. B.* (1996), Semantic types of measure [Semanticheskie tipy razmera], Moscow Linguistic Journal [Moskovskij lingvisticheskij zhurnal], Vol. 2. RSUH, Moscow, pp. 80–96.
7. *Borschev V. B.* (2014a), On the integration of formal and lexical semantics [Ob integracii formal'noj i leksicheskoj semantiki], In Formal Approaches to Russian Linguistics, MGU, http://otipl.philol.msu.ru/library/seminars/farl/handouts.php
8. *Borschev V. B.* (2014b), Once more on types and sorts (or on the integration of formal and lexical semantics) [Eshche paz o tipax i sortax (ili ob integracii formal'noj i leksicheskoj semantiki)]. In Language. Constants. Variables: To the memory of Alexander Evgenevich Kibrik [Jazyk. Konstanty. Peremennye: pamjati Aleksandra Evgen'evicha Kibrika], Aleteja, Saint Petersburg, pp. 38–56.
9. *Borschev V. B., Partee B. H.* (1999a), Semantics of genitive construction: different approaches to formalization [Semantika genitivnoj konstrukcii: raznye podxody k formalizacii]. In Typology and Linguistic Theory: From Description to Explanation. For the 60th birthday of Aleksandr E. Kibrik [Tipologija i teorija jazyka: Ot opisanija k ob"jasneniju. K 60-letiju Aleksandra Evgen'evia Kibrika], Jazyki Russkoj Kul'tury, Moscow, pp. 159–172.
10. *Borschev V. B., Partee B. H.* (1999b), Semantic types and the Russian genitive modifier construction, In Formal Approaches to Slavic Linguistics: The Seattle Meeting 1998, Michigan Slavic Publications, Ann Arbor, pp. 39–57.

11. *Borschev V. B., Partee B. H.* (2001), Genitive modifiers, sorts, and metonymy, Nordic Journal of Linguistics, Vol. 24, pp. 140–160.
12. *Borschev V. B., Partee B. H.* (2001), Ontology and metonymy. In Ontology-Based Interpretation of Noun Phrases. Proceedings of the First International Onto-Query Workshop, Department of Business Communication and information Science, University of Southern Denmark, Kolding, pp. 121–138.
13. *Borschev V. B., Partee B. H.* (2004), Genitives, types, and sorts: The Russian genitive of measure. In Possessives and Beyond: Semantics and Syntax (UMOP 29), GLSA Publications, Amherst, MA, pp. 29–43.
14. *Borschev V. B., Partee B. H.* (2011), The genitive of measure in Russian, Types and sorts [Genitiv mery v russkom jazyke, tipy i sorta], In Word and Language. A Collection of Articles for the 80th Birthday of Academician Ju. D. Apresjan [Slovo i jazyk. Sbornik statej k 80-letiju akademika Ju. D. Apresjana], Jazyki slavjanskix kultur, Moscow, pp. 95–137.
15. *Chierchia G.* (1982), Bare plurals, mass nouns, and nominalization, In Proceedings of the First West Coast Conference on Formal Linguistics, Stanford Linguistics Association, Stanford University, Stanford
16. *Chierchia G.* (1998a), Plurality of mass nouns and the notion of "semantic parameter", In Events and Grammar, Kluwer, Dordrecht, pp. 53–103.
17. *Chierchia G.* (1998b), Reference to kinds across languages, Natural Language Semantics, pp. 339–405.
18. *Dowty D.* (1979), Word meaning and Montague grammar. The semantics of verbs and times in Generative Semantics and in Montague's PTQ, Synthese Language Library, Reidel, Dordrecht.
19. *Kamp H., Partee B.* (1995), Prototype theory and compositionality, Cognition, Vol. 57, pp. 129–191.
20. *Krifka M.* (1995), Common nouns: a contrastive analysis of Chinese and English, In The Generic Book, The University of Chicago Press, Chicago, pp. 383–411.
21. *Landman F.* (2004), Indefinites and the Type of Sets, Explorations in Semantics, Blackwell Publishing, Oxford.
22. *Link G.* (1983), The logical analysis of plurals and mass terms: A lattice-theoretical approach, In Meaning, use and the interpretation of language, Walter de Gruyter; Berlin, New York, pp. 303–323. Reprinted in Portner and Partee (2002), pp. 127–146. Reprinted in Link G. (1998), Algebraic Semantics in Language and Philosophy: CSLI lecture notes No. 74, CSLI Publications, Stanford, Calif., pp. 11–34.
23. *Parsons T.* (1970), An Analysis of Mass and Amount Terms, Foundations of Language, Vol. 6, pp.363–385.
24. *Partee B. H., Borschev V.* (2010). Bare 'milk' in 'glass of milk' in English and Russian, Paper presented at Workshop on Bare Noun Phrases, Bar Ilan University, Israel.
25. *Partee B. H., Borschev V.* (2012a), Sortal, relational, and functional interpretations of nouns and Russian container constructions, Journal of Semantics, Vol. 29, pp. 445–486.
26. *Partee B. H., Borschev V.* (2012b), Dva stakana moloka: substances and containers in genitive of measure constructions in Russian, Russian Language and Linguistic Theory [Russsskij jazyk v nauchnom osveshchenii] Vol. 2 (24), pp. 140–166.

27. *Pustejovsky J.* (1995), The Generative Lexicon, The MIT Press.
28. *Rothstein S.* (2009a), Measuring and counting in Modern Hebrew, Brill's Annual of Afroasiatic Languages and Linguistics 1, pp. 106–145.
29. *Rothstein S.* (2009b), Individuating and measure readings of classifier constructions: Evidence from Modern Hebrew, Ms., Handout for a Conference "Atoms and Laws of the NP", Utrecht University. Utrecht.
30. *Schwarzschild R.* (2002), The grammar of measurement, Proceedings of Semantics and Linguistic Theory XII, Ithaca, NY, pp. 225–245.
31. *Schwarzschild R.* (2005), Measure phrases as modifiers of adjectives, Recherches Linguistiques de Vincennes 34: L'adjectif, Paris, pp. 207–228.
32. *Schwarzschild R.* (2006), The role of dimensions in the syntax of noun phrases, Syntax Vol. 9/1, pp. 67–110.
33. *Thomason R.* (1972), A Semantic Theory of Sortal Incorrectness, Journal of Philosophical Logic 6 Vol. 1, pp. 209–258.

ВИРТУАЛЬНЫЙ РУССКИЙ КОРПУС С СЕМАНТИЧЕСКОЙ РАЗМЕТКОЙ И ПОИСК ДЕФЕКТОВ В СЛОВАРЕ-ПОСРЕДНИКЕ

Диконов В. Г. (dikonov@iitp.ru)
ИППИ РАН, Москва, Россия

Порицкий В. В. (v.poritski@gmail.com)
БГУ, Минск, Беларусь

Ключевые слова: векторное пространство, корпус, семантическая разметка, лексические ресурсы, словарь-посредник, семантика

A VIRTUAL RUSSIAN SENSE TAGGED CORPUS AND CATCHING ERRORS IN A RUSSIAN ↔ SEMANTIC PIVOT DICTIONARY

Dikonov V. G. (dikonov@iitp.ru)
IITP RAS, Moscow, Russia

Poritski V. V. (v.poritski@gmail.com)
BSU, Minsk, Belarus

There are areas in computational linguistics, where a word-sense tagged corpus becomes a necessary prerequisite or gives a significant boost to research. Unfortunately, publicly available corpora of this kind are extremely rare and making them from scratch is a very long and costly process. No corpus of Russian with unambiguous word-sense tags has been published so far. This paper describes an experimental approach of creating a virtual equivalent of a Russian sense tagged corpus and putting it to some real use. The virtual corpus was created using two public resources: the English SemCor corpus and our free multilingual semantic pivot dictionary, called the "Universal Dictionary of Concepts". The dictionary provides information sufficient to find sense-specific translations for nearly all sense-tagged words in SemCor. However, the pivot dictionary itself is under development and we are looking for the ways to improve it. We used the existing Russian volume of the pivot dictionary to calculate lexical context vectors for individual senses of 13,832 Russian words, supposedly equivalent to the

vectors that could be obtained from a real Russian translation of SemCor. Another set of vectors representing real usage of the same Russian words was extracted from a medium-size corpus of Russian without any semantic markup. The vector similarity score proved to be a useful factor in judging the correctness of links between Russian words and word senses similar to ones registered in the Princeton Wordnet. It helped to rank over 21,000 of such links out of 56,000 known and significantly reduce the amount of the manual work required to proofread the dictionary.

Keywords: vector space, corpus, sense tagging, lexical resources, pivot, semantics, dictionary

1. Introduction

The motive for using the approach outlined in the abstract was absence of a free Russian corpus with semantic markup and scarcity of publicly available lexical and semantic resources that would formally describe the meanings of Russian words in a machine-readable form. One of the authors has already invested some effort in plugging the latter gap while developing an open multilingual semantic resource "Universal Dictionary of Concepts", described in [Dikonov, 2013], [Boguslavsky, Dikonov, 2009]. It is further referred to as pivot dictionary. The goal of the work presented here is two-fold. Firstly, we try to ensure good quality of the pivot dictionary by correcting most of the eventual errors. Secondly, we do what is possible to prepare raw data that could be used for developing and improving Russian word sense disambiguation tools (rule-based and statistical models). We shall briefly describe the resources and devices we used in sections 2–5, explain the process of result evaluation in section 6 and present the results in section 7.

2. Pivot dictionary

The Universal Dictionary of Concepts is a repository of fine-grained semantic concepts, which are equivalent to word senses, with translations into several natural languages. The senses are organized into semantic classes, supported by SUMO ontology [Pease, 2011], and linked by a network of semantic relations. The dictionary serves as a lexicon of an artificial computer interlingua called UNL and uses UNL "Universal Words" as unique identifiers of the senses. Its structure makes it a good neutral semantic pivot dictionary, which is not limited to the lexicon of any single natural language.

The first versions were bootstrapped by integrating available free lexical and ontological data, including Princeton Wordnet, by various automatic methods. However, further development along the lines of simple data merging was hampered by the fact that every imported error in the links between words and abstract concepts tends to multiply and produce even more entropy, as soon as already known links get used to classify new data.

Currently the Russian part of the dictionary covers about 33,000 entries, not counting most proper names and multiword terminology. These words are linked to over 56,000 senses, including some specific to the Russian language. The target

quality level at the early automatic data acquisition phase was set at no less than 90% of correct Russian word↔sense links. A lot of work has already been done to improve it by proofreading critical parts of the dictionary. As a result, the estimated percentage of wrong links decreased to approx. 2.5% with about 3% of additional questionable or vague translations[1], such as *гореть (burn)* instead of *полыхать (flare, burn up)* in the sense "Burn brightly", which is a hyponym of *burn* "Undergo combustion". The remaining errors were randomly scattered among more than 56,000 word↔sense pairs. We needed a way to concentrate the errors in a smaller section of the data to reduce labor intensive straightforward editing. One possibility to do it is to employ a vector space model and a corpus.

3. Virtual corpus

Our source of sense-tagged text was SemCor [Mihalcea, 1998]—a subset of Brown corpus with manual tagging by Wordnet 2.1 senses. It was supplemented by SensEval [Kilgarriff, 1998] 2 and 3 benchmark files converted to SemCor format. This gave us 37,698 English sentences containing 724,207 words.

In 2010 we evaluated the potential of sense tags in SemCor to improve the quality of syntactic parsing and made dependency trees for 37,136 (98.5%) sentences in SemCor+SensEval [Dikonov, D'jachenko, 2010]. We used an experimental build of the ETAP-3 parser, which was modified to use external tagging, either manual or from another parser. The use of semantic annotation helped us to build better syntactic trees.

The tree-tagged SemCor+SensEval corpus contains both the original sense tagging and extra tags given to words, which can have only one meaning according to the pivot dictionary. The total number of sense tagged instances of English words in our corpus is 144,723 (21,978 unique senses) and they make 782,009 unique pairs. Unfortunately, large portions of SemCor have very sparse annotation, e.g. only verbs are disambiguated in 166 files out of 352. Sentences with only one tagged word or excessive linear distance between tags are useless for measuring co-occurrence of senses. We used both linear window sized 1, 2 or 3 words on two sides and syntactic dependencies to learn co-occurrence statistics for pairs of senses. The largest set of such pairs was built by combining dependencies with a 3-word window. This option was used for all subsequent steps.

Our pivot dictionary provides sense-tagged words in the corpus with sense-specific translations into Russian. For example, the English noun *bill* has many senses and each is translated into Russian in a different way:

- "A sign posted in a public place"—*афиша*
- "A statute in draft before it becomes law "—*законопроект*
- "A statement of money owed for goods shipped or services rendered"—*счет*
- "A piece of paper money"—*купюра, банкнота*
- "A list of particulars "—*список*
- "A male given name "—*Билл*

[1] This estimation was done by taking random samples of 200 links and counting defects found during proofreading of the samples.

- "A brim that projects to the front to shade the eyes"—*козырек*
- "Horny projecting mouth of a bird "—*клюв.*

In figure 1 below it is translated as *законопроект* because the corpus has a tag indicating that *bill* means "A statute in draft before it becomes law". Almost all sense-tagged words receive one or more Russian equivalents is this way. Russian translation equivalents which are multiword phrases are converted into sequences of independent lemmas.

From such data we can compute mutual co-occurrence frequencies of Russian words which should be close enough to the frequencies, that would be observed in a real Russian translation of the same text. The numbers would never actually match for many reasons. One of them is that some words in SemCor lack sense tags, like the words *school*, *student*, *reduce* in figure 1. Another is that the pivot dictionary is incomplete and some words remain untranslated. Nevertheless, at this step we obtain potentially useful data approximating the data that could be obtained from a non-existent Russian SemCor-like corpus.

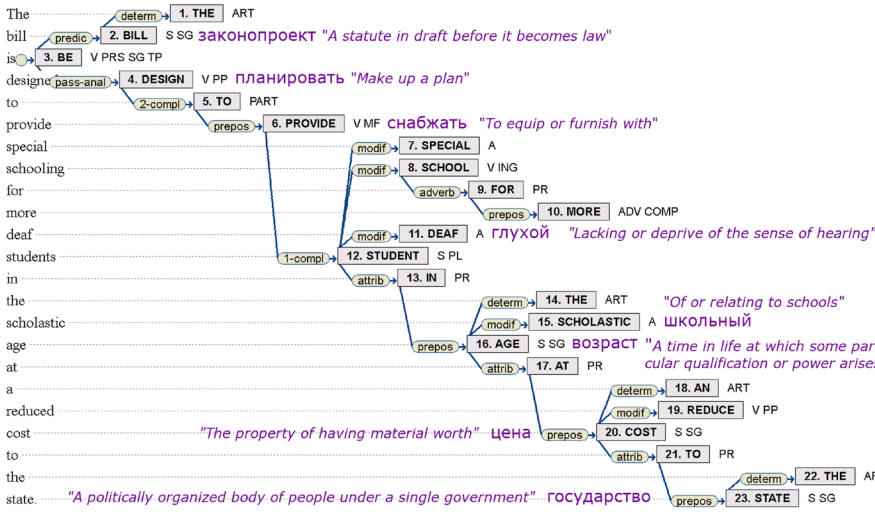


Fig. 1: A sample of SemCor data with extra annotation: dependency tree produced by ETAP3 automatic parser, guided by semantic tags, and Russian translations of semantically tagged words

4. Semantic vectors

The set of numbers associated with a word sense and showing, how many times different other words occur in the context of the word used to express that sense, makes up a numerical vector. The vectors for different senses of the same word are different, because the context neighbors of the senses usually differ. We call such vectors semantic, as opposed to lexical vectors associated with non-disambiguated words.

Predicted semantic vectors are sensitive to defects of the pivot dictionary used to produce them. It is possible to identify vectors based on wrong translations by comparing them to other vectors built from some benchmark data and representing correct use of the Russian words. Ideally, the benchmark should provide semantic vectors representing the same word senses, but in our case this was not possible.

However, the semantic vectors produced from the virtual corpus can still be compared with lexical vectors representing all available contexts of Russian translations for the semantic vector's base sense. Any false translations in the semantic vectors reduce similarity with the benchmark lexical vectors. A false translation of the vector base sense causes the comparison to be made with entirely different set of contexts, belonging to the word which does not have this sense. This results in a very noticeable difference. For example, one of the real detected errors was that the sense *weld(icl>join>do,agt>thing,obj>thing) "Join together by heating*" was wrongly linked to the Russian word *сплачивать*, which means "Unite closely or intimately" *weld(icl>unite>do,cob>thing,agt>volitional_thing,obj>thing)*. The first sense is likely to be found in phrases like *сваривать панели (weld panels)* but the wrong translation meant that the virtual corpus offered **сплачивать панели* instead. The latter phrase never occurs in real Russian texts and the corresponding semantic vector fails comparison with the benchmark vector of the word *сплачивать*.

A bunch of correct semantic vectors, representing all senses of some word, put together is likely to show close semblance to the benchmark lexical vector of the word. Our hypothesis is that a single correct semantic vector still has enough similarity with its benchmark to be distinguishable from random errors.

5. Benchmark corpus

We used a benchmark Russian corpus of approximately 17 mln tokens. The corpus contains samples of present-day Russian fiction (10 mln tokens) and newspaper articles (the rest). To obtain lemmas, we merged the output of MyStem [Segalovich 2003] and TreeTagger with Russian parameter file [Schmid 1992; Sharoff et al. 2008]. In most cases TreeTagger works as a disambiguator over the output of MyStem, but its lexical coverage is rather narrow, since the parameter file has been trained on the disambiguated portion of Russian National Corpus. For some of the wordforms not recognized by TreeTagger, MyStem produces a unique lemma, so that a simple fallback strategy is available. MyStem also helps to deduce lemmas for several trickier classes of tokens: compound nouns, age designations like *23-летний* (23 years old) etc.

The key idea was to design a high-dimensional vector space, such that both senses and their purported Russian equivalents could be represented as points thereof. The basis of this space was made up, rather straightforwardly, of lemmas attested both in the virtual sense-tagged Russian corpus and in the benchmark corpus. This amounts to ca. 10^4 distinct lemmas. To compute a suitable similarity score between two 10^4-dimensional vectors is a tractable task, so no further dimensionality reduction was done. Note however, that moderate size of the basis brings not only

computational ease, but also scalability issues: no matter how large the benchmark corpus is, most of the co-occurrence statistics collected in it will remain unused.

Four sets of benchmark vectors have been computed, with symmetric linear context window size ranging from 1 to 4 tokens. We used cosine similarity which had performed well in earlier experiments on coarse-grained synonym identification [Poritski, Volchek 2013]. Pairwise similarity computations were run on raw co-occurrence counts as well as on PMI weighted vectors (for the definition of PMI weighting scheme see, e.g., [Manning, Schütze 2003, p. 178]). The similarity score value are numbers between 1 and 0. The value of 1 is given to pairs of vectors which are elementwise proportional (high similarity). Absence of similarity (totally different vectors) is marked with 0 (for raw frequency counts) or −1 (with PMI applied). However negative scores under PMI weighting turned out to be quite rare and were counted as zeros

As a result we built several versions of similarity scores, using different ways of finding word pairs and calculating vector similarity. Each version was presented as a table containing three columns: a Russian word, pivot word sense designation (UNL universal word) and the similarity score of the semantic and benchmark vectors, as shown in figure 2. The number of lines was 22,874, which corresponds to the number of word senses occurring in SemCor+SensEval and translatable through the pivot dictionary.

беспокоить	bother(icl>trouble> cause>do,agt>thing,obj> person,met>uw)	0.756537995710645	*To cause inconvenience or discomfort to*
камень	stone(icl>material> thing,equ>rock)	0.0501434171547867	*Material consisting of the aggregate of minerals*
камень	stone(icl>natural_ object>thing,equ>rock)	0.0438500966889403	*A lump or mass of hard consolidated mineral matter*
~~гореть~~ → полыхать	burn_up(icl>burn>occur, equ>flare,obj>thing)	0.0178584089774577	*Burn brightly*
~~сплачивать~~ → сваривать	weld(icl>join>do,agt> thing,obj>thing)	0	*Join together by heating*

Fig. 2. Lines from the similarity score table with comments and corrections

6. Evaluation

The links in the similarity tables were sorted by decreasing of the similarity score. At this point we needed to find, which word↔sense links were wrong. It was done in several iterations.

At first, one of the tables was deemed most promising by comparing the score and position of a couple of already known errors and subjected to selective manual examination. All bad links found were marked as errors or overly vague translations by different symbols. Samples of 200 lines were taken from different parts of the table,

starting from lines 0 (highest scores, 1 error), 8,000 (scores of ~0.02, 3 errors), 10,100 (scores of ~0.009, 13 errors), 12,127 (scores of ~0.001, 7 errors) and 18,600 (zero similarity scores, 16 errors). This first attempt produced a test set of 63 defects (40 errors and 23 vague translations). It also showed that the probability of errors in word↔sense pairs increased as vector similarity score dropped and the concentration of errors in different parts of the table changed from 0.5% to 8% per 200 line sample.

This allowed us to do a better numerical estimation and choose another table, which seemed more likely to have the optimum parameters. We tried to find a threshold in similarity score. Again, 200 line samples were taken starting from lines 4,800 (scores of ~0.05, 8 errors), 6,800 (scores of ~0.04, 8 errors), 8,500 (scores of ~0.03, 10 errors), 10,000 (scores of ~0.024, 17 errors), 11,800 (scores of ~0.015, 10 errors) and further 580 lines with zero score (46 errors). This time the overall error distribution curve, similar to ones shown in figure 3, actually got flatter and some fluctuations became visible.

The same word↔sense links receive different scores in tables built with different settings, so the errors found in the first chosen table were scattered around the second table randomly. Combining the error sets produced a more evenly distributed test set. A review of the combined set confirmed that there are certain other factors helpful in selecting likely errors. Our pivot dictionary allows to differentiate between polysemic and monosemic words in all supported languages, including Russian. It also has technical flags showing the amount of attention given to each word↔sense link. It is rather obvious that polysemic words and less reviewed links are more suspicious and our data confirmed it. As a result, further samples were gathered and all non-reviewed polysemic words occurring in the zero similarity zone were checked. This gave us a test set of 1,141 bad links (659 errors and 512 vague translations).

7. Results

The following two diagrams show the distribution of all defects discovered until now in the cosine similarity score tables, as shown in figure 2. Figure 3 illustrates the dependency between the score (X) and the number of defects (Y) in one of the best ranking tables which was computed with the following settings:

- co-occurrence of senses in the virtual corpus within linear window of width 3 and within the range of syntactic dependencies;
- co-occurrence of words in the benchmark corpus within linear window of width 2;
- PMI weighting applied.

Higher score value means better alignment between semantic and benchmark vectors. The first defects start to appear when the score drops below 0.16. The three colored bars correspond to all defects, errors properly and vague translations.

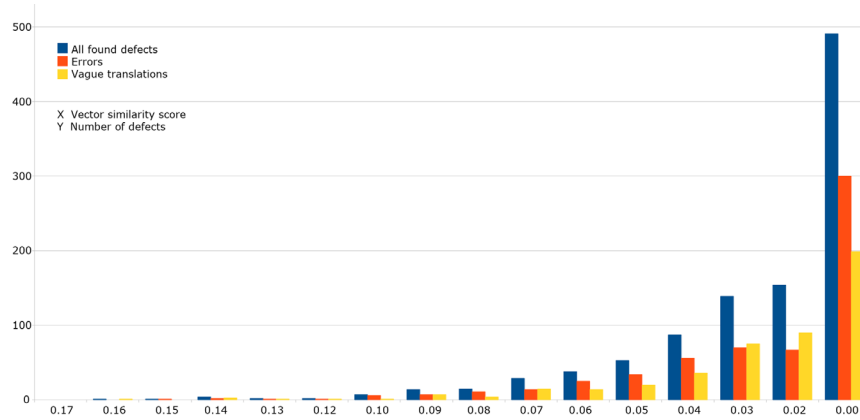


Fig. 3. Distribution of defects according to the similarity score

The diagram in figure 4 shows, how the number of defects per 1,000 lines changes from beginning to the end in several versions of the table, all ranked by cosine similarity score in decreasing order. Each version represents a different combination of options used to produce the benchmark vectors and is shown by a different curve in the diagram. The possible options are:

- linear window width (1–4) in the benchmark corpus;
- PMI weighting (yes/no);
- frequency count strategy for vector elements, which are known to be synonyms (sum all / take maximum).

Here PMI weighting is always on, because it makes the result consistently better. The horizontal line shows calculated average level, when all defects are scattered randomly.

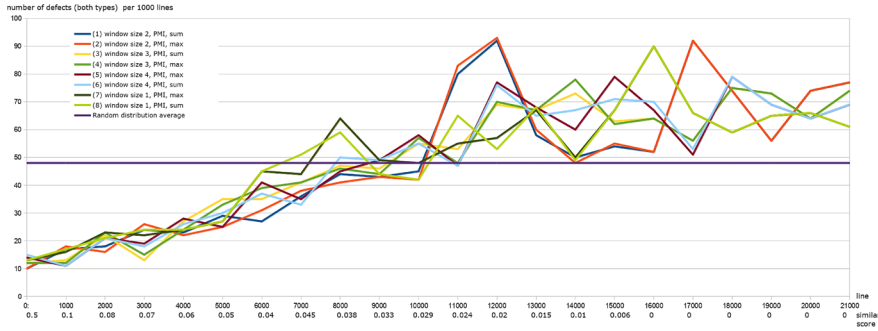


Fig. 4. Distribution of defects by line number in the similarity score tables

Already reviewed data confirms previous quality estimations of the pivot dictionary. Before this work, the predicted total number of errors in its Russian part ran

between 2,070 and 2,275 out of 56,000 links. The same number for vague translations was about 2,500. Defects already exposed by the procedures described in section 6 constitute approximately 29% of the predicted total number of errors and about 20% of vague translations. At the same time, the reviewed portion of the vector similarity table at the time of writing (6,218 links) is only 10.6% of the total contents of current Russian dictionary.

8. Conclusion

The virtual corpus proved to be a reasonably useful tool for narrowing down the search for anomalies in relations between Russian words and pivot word senses. It can make the process of discovering and fixing dictionary defects almost 3 times faster than baseline. This is a sound practical outcome.

Although the two sets of vectors are based on different things—individual word senses and non-disambiguated words—comparing them was fruitful. It is possible to merge semantic vectors representing all registered senses of the same word to obtain a predicted lexical vector. That would make a completely fair "apples to apples" comparison. It has not been done because we cannot state at this point that all words in our pivot and Wordnet have complete description of their polysemy. We may do it at a later stage to facilitate the search of Russian words, which lack certain key senses in our pivot dictionary.

The amount of publicly available sense disambiguated corpus data is dismally limited for English and is simply zero for Russian. There are some Russian resources though, which are not public in the sense that they cannot be freely downloaded and used. One example is the semantic annotation layer within the Russian National Corpus (RNC) [Lashevskaja, Shemanaeva 2008]. It is different from the SemCor data used in this work in several respects. RNC does not label individual instances of words with any concrete word senses. Instead, they receive a set of taxonomic, mereological and derivational tags, assigned by software according to the RNC's internal semantic dictionary. Unlike SemCor, no manual disambiguation has been done in RNC. The tags, however, were filtered with manually formulated rules to remove tags violating known contextual restrictions. The resulting partially disambiguating semantic markup is used by the online RNC search engine.

Even if a real manually sense-tagged Russian corpus will be developed, we can hardly expect it to be larger than SemCor. It is possible to improve the situation by supplementing one small corpus with another small corpus made for a different language. It requires a reliable pivot, which allows to match or relate different sets of word sense labels. Such supplementing may work for projects that generalize the senses to a coarser grain level or rely on statistics to smooth over small problems. The current version of the pivot dictionary is available for download from the git repository at https://github.com/dikonov/Universal-Dictionary-of-Concepts. The tree-tagged virtual corpus files with Russian translations of the sense tagged English words will be published when the process of proofreading the links will be near completion. Snapshots can be found at https://github.com/dikonov/SemCorRus.

References

1. *Boguslavsky I., Dikonov V.* (2009), Universal Dictionary of Concepts [Universal'nyj slovar' konceptov], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2009" [Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii "Dialog 2009"], Bekasovo, pp. 91–96.
2. *Dikonov V., Boguslavsky I.* (2009), Semantic Network of the UNL Dictionary of Concepts, Proceedings of the SENSE Workshop on Conceptual Structures for Extracting Natural Language Semantics, Moscow, available at: http://ceur-ws.org/Vol-476/paper2.pdf.
3. *Dikonov V., D'jachenko P.* (2010), An Experiment in Automatic Building of English Dependency Trees Governed by Externally Provided Incomplete Tagging [Eksperiment po postroeniju sintaksicheskoj struktury anglijskih predlozhenij s ispol'zovaniem zaranee izvestnyh fragmentarnyh dannyh], Proceedings of ITaS'10, Gelendzhik, pp. 310–319.
4. *Kilgarriff A.* (1998), SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs, Proceedings of LREC'98, Granada, pp. 581–588.
5. *Lashevskaja O. N., Shemanaeva O. Yu.* (2008), Semantic Annotation Layer in Russian National Corpus: Lexical Classes of Nouns and Adjectives, Proceedings of LREC'08, Marrakech, pp. 3355–3358.
6. *Manning C. D., Schütze H.* (1999), Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA.
7. *Mihalcea, R.* (1998), SemCor semantically tagged corpus, SenseEval 2 & 3 data in SemCor format. http://www.cse.unt.edu/~rada/downloads.html
8. *Pease A.* (2011), Ontology: A Practical Guide, Articulate Software Press, Angwin, CA.
9. *Poritski V. V., Volchek O. A.* (2013), Building a Vector Space Model of Meaning for Russian: A Preliminary Study [Postroenie vektornoj semanticheskoj modeli na osnove russkojazychnyh tekstov: pervye eksperimenty], Proceedings of ITaS'13, Svetlogorsk, pp. 114–119.
10. *Schmid H.* (1992), Probabilistic Part-of-Speech Tagging Using Decision Trees, Proceedings of International Conference on New Methods in Language Processing, Manchester, pp. 44–49.
11. *Segalovich I.* (2003), A Fast Morphological Algorithm With Unknown Word Guessing Induced by a Dictionary for a Web Search Engine, Proceedings of MLMTA'03, Las Vegas, pp. 273–280.
12. *Sharoff S., Kopotev M., Erjavec T., Feldman A., Divjak D.* (2008), Designing and Evaluating a Russian Tagset, Proceedings of LREC'08, Marrakech, pp. 279–285.

ДИСКУРСИВНЫЕ СЛОВА В ОБЩЕВОПРОСИТЕЛЬНЫХ ПРЕДЛОЖЕНИЯХ: РУССКО-НЕМЕЦКИЕ СООТВЕТСТВИЯ[1]

Добровольский Д. О. (dm-dbrv@yandex.ru),
Левонтина И. Б. (irina.levontina@mail.ru)

Институт русского языка
им. В. В. Виноградова РАН, Москва, Россия

В докладе рассмотрены русские частицы *разве, неужели, что, что ли, как* и др. *(Ты что / что ли / как, с нами идешь?)* и их наиболее близкий немецкий эквивалент *etwa (Gehst du etwa mit?)*. Материал частиц позволяет ясно увидеть разницу между переводимостью и семантическим тождеством. При богатом репертуаре вопросительных частиц и в русском, и в немецком языках для каждого высказывания можно подобрать хороший эквивалент, причем с частицей. Однако из этого не следует, что сами частицы семантически эквивалентны. Анализ показывает, что смысловое сходство между самими частицами довольно отдаленное.

Ключевые слова: дискурсивные слова, частицы, общевопросительные предложения, семантика, прагматика, русский язык, немецкий язык

DISCOURSE WORDS IN GENERAL QUESTIONS: RUSSIAN-GERMAN NEAR-EQUIVALENTS

Dobrovol'skij D. O. (dm-dbrv@yandex.ru),
Levontina I. B. (irina.levontina@mail.ru)

Vinogradov Russian Language Institute
of the Russian Academy of Sciences, Moscow, Russia

The paper discusses Russian discourse words such as *razve, neuzheli, chto, chto li, kak,* etc. Cf. *Ty chto / chto li / kak, s nami idesh'*? ≈ 'What about you, are you coming along?', and their German near-equivalent etwa (cf. *Gehst*

[1] Работа выполнена при поддержке РФФИ (грант 13-06-00403) и Программы фундаментальных исследований Президиума РАН «Корпусная лингвистика».

du etwa mit?). Our data show that translatability and semantic equivalence are different phenomena. Both Russian and German possess a rich inventory of question particles, which makes it possible to find a suitable translation for nearly every utterance, even a translation containing a particle. However, this does not imply that the corresponding particles are semantically equivalent. The analysis shows that such particles, being functionally equivalent, i.e. interchangeable in particular utterances, display rather remote semantic resemblance. The German particle *etwa* is conceptually based on the idea of approximateness. That is why it weakens the illocutionary force of the utterance, whereas the Russian particles *chto, chto li, kak* directly appeal to the interlocutor and, therefore, reinforce the speaker's attitude. However, both German *etwa* and Russian *chto, chto li, kak* stress the speaker's involvement in the situation. This property determines their functional similarity.

Key words: discourse words, particles, general questions, semantics, pragmatics, Russian, German

Среди дискурсивных слов, представленных в русском и немецком языках, особое место занимают единицы, связанные с идеей побуждения в широком смысле, в том числе и вопроса. Это естественно: ведь говорящий вступает в определенные отношения с адресатом, часто усложняя при этом иллокутивный акт. Поскольку говорящему нужно получить какую-то реакцию (вербальную и невербальную), то вступают в действие изощренные стратегии. Они могут быть разные: надавить на адресата, разжалобить его, затемнить свои цели. Мы уже писали об интересных русско-немецких соответствиях между побудительными частицами [Добровольский, Левонтина 2010]. В данной работе мы обращаемся к дискурсивным средствам, специфичным для общевопросительных предложений. О различных иллокутивных эффектах в вопросительных предложениях см. [Булыгина, Шмелев 1987]; ср. также [Рестан 1972].

Лучше всего изучены частицы *разве* и *неужели*. Наиболее точно и подробно, на наш взгляд, эти две частицы описаны в замечательной работе Т. В. Булыгиной и А. Д. Шмелева [Булыгина, Шмелев 1997: 270–281]: «Общим для этих частиц является то, что они маркируют вопрос-реакцию. Это значит, что вопросы с этими частицами не могут задаваться «ни с того ни с сего», просто потому, что у говорящего возникла соответствующая информационная потребность. Высказывания с частицами *разве* и *неужели* уместны лишь в тех случаях, когда в поле зрения говорящего попала ситуация, не соответствующая его ожиданиям» [Булыгина, Шмелев 1997: 270]. В указанной работе подробно разобраны различия между *разве* и *неужели*.

Стоит сравнить и другие дискурсивные слова для общевопросительных предложений (см. подробнее [Левонтина, в печати]). Рассмотрим частицу *что (чего, чё)*:

Ты ***что****, с нами идешь?*
— Ее муж… — А она ***что****, замужем?*
Ты ***что****, издеваешься?*

Интересно, что, в отличие от большинства частиц, которые тяготеют к клитичности, безударности, *что*, напротив, полноударно и часто даже акцентно выделено. В письменной передаче этому просодическому свойству соответствует постановка запятой после *что*.

На первый взгляд кажется, что по значению *что* очень похоже на *разве*. Действительно, в наши первые два примера легко подставить *разве*:

*Ты **разве** с нами идешь?*
*— Ее муж… — А она **разве** замужем?*

Однако в третий пример подставить *разве* невозможно — во всяком случае, без существенного изменения смысла, ср. прагматически странное:

*Ты **разве** издеваешься?*

В следующей паре обе фразы возможны, но они оказываются почти антонимичными:

*Он **что**, издевается?*
*Он **разве** издевается?*

В первом случае говорящий готов счесть, что *издевается,* во втором это кажется говорящему малоправдоподобным. Этот пример показывает, что *что* и *разве* подразумевают разные установки говорящего.

В случае *что* говорящий воспринял какую-то информацию, которая заставляет его счесть, что имеет место ситуация Р. Он не ожидал, что Р, но склонен, хотя и с некоторым усилием, признать, что Р имеет место.

В случае *разве* говорящий воспринял какую-то информацию, которая заставляет его счесть, что имеет место ситуация Р. Он не ожидал, что Р (до этого места толкование совпадает), и не готов сразу признать, что Р имеет место.

В некоторых случаях это различие оказывается прагматически несущественным, тогда частицы взаимозаменимы, как в примере *— Ее муж… — А она **что, / разве** замужем?* Кроме того, *что* и *разве* могут выступать и совместно: *— А она **что, разве** замужем?*

Совершенно естественно, что *что* не заменяется на *разве* в ситуации, когда подлинная цель говорящего состоит не в получении информации, а выражении оценки — говорящий риторически предлагает малоправдоподобную интерпретацию ситуации, чтобы показать, что ситуация, с его точки зрения, совершенно ненормальна: *Ты что, издеваешься / смеешься / шутишь / с дуба рухнул / совсем спятил?; Это что, шутка?*

На частицу *что* похожа частица *что ли* (эти две частицы также могут употребляться совместно):

Ты с нами ***что ли*** *идешь?*
Ты издеваешься ***что ли****?*
А она ***что****, замужем* ***что ли****?*

Тем не менее, *что* и *что ли* не полностью синонимичны. Рассмотрим следующий пример:

— Подай тарелку. — Эту ***что ли****?*

В этом случае замена на *что (****Что****, эту?)* нежелательна или во всяком случае существенно изменит смысл высказывания. Дело в том, что *что ли*, в отличие от *что*, скорее подразумевает, что у говорящего, в общем-то, не было каких-либо специальных ожиданий по поводу ситуации Р, он получил информацию, она для него просто новая, поэтому он хочет удостовериться, что правильно ее воспринял. Или, возможно, она даже не новая, а неточная или позабытая: *Какой дом — пятый,* ***что ли****?; Где поворачивать, здесь,* ***что ли****?* Естественно, *что* здесь неуместно.

Из сказанного выше следует, что с частицей *разве* частица *что ли* не должна хорошо сочетаться (они выражают противоположные установки говорящего в отношении готовности принять новую информацию). И действительно, эти две частицы не выступают совместно:

А она* ***разве* *замужем* ***что ли****?*

Следует обратить внимание на такую единицу, как *ну что*, которая, несмотря на кажущееся сходство с *что* и *что ли*, сильно от них отличается.

Так, во фразе *Ты с нами* ***что ли*** *идешь?* вполне возможна замена:

Ну что*, ты с нами идешь?*

Правда, *Ты издеваешься* ***что ли****?* едва ли синонимично ***Ну что****, ты издеваешься?*

Но главное различие не в этом. Дело в том, что *ну что* не специфично для вопросительных контекстов. *Ну что* предполагает, что своей репликой говорящий подводит итог предшествующей дискуссии, размышлениям, наблюдениям или ожиданию. И в данном случае не очень существенно, содержит ли реплика вопрос, сообщение или побуждение; ср.:

Ну что*, я поехал;*
Ну что*, давай звони.*

Как мы видим, *ну что* может использоваться и в вопросительных предложениях, но сама эта единица никакой вопросительности не выражает.

К рассматриваемой группе вопросительных частиц примыкает еще слово *как*:

Ты ***как****, с нами идешь?*

Существенно, что в другой нашей фразе *как* невозможно:

Ты **как, издеваешься?*

Как, подобно ч*то ли*, подразумевает, что говорящий не имел заранее никаких ожиданий. Но если ч*то ли* означает, что говорящий узнал что-то, что заставляет его подозревать Р, *как* не указывает на получение говорящим какой-то новой информации. Для него существует альтернатива, Р или не Р, и он хочет узнать, какой из двух вариантов имеет место. Поэтому вполне естественно:

*Ты **как**, с нами идешь или нет <или дома остаешься>?*

Ни ч*то*, ни ч*то ли* в данной фразе невозможны.

Указанные семантические особенности частицы *как* по сравнению с рассмотренными ранее определяют важное функциональное отличие этой единицы от остальных. Как мы помним, конституирующим свойством таких слов, как *разве* и *неужели*, а также наших ч*то* и ч*то ли* является входящая в их значение идея конфликта между представлениями, имевшимися у говорящего до момента речи, и информацией, полученной им непосредственно перед этим моментом. По этой причине данные единицы, являясь вопросительными, а значит, подразумевающими ответную реплику, в то же время сами чаще всего формируют ответную реплику (ведь информация, полученная непосредственно перед моментом речи, скорее всего, содержалась в предшествующей реплике собеседника). Что же касается частицы *как*, она такого конфликта не подразумевает — и совершенно закономерным образом не обязана использоваться в ответных репликах. Более того, такое ее использование затруднено:

*— Ее муж… — *А она **как**, замужем?*

Такое возможно, только если вторая реплика не является собственно ответом на первую, а связана с нею по смыслу более опосредованно:

*— Хочешь с ней познакомиться? — А она **как**, замужем?*

Второй говорящий не отвечает на поставленный вопрос, а хочет для ответа на него получить дополнительные сведения.

Еще одну группу вопросительных частиц образуют сочетания с *не: случайно (не), случаем (не), часом (не):*

*А это **случайно** не пятый дом?*
*У него **случайно** нет брата?*
*А ты **случаем/часом** не шпион?*

Как и вообще вопросы с отрицанием (ср. *У тебя нет ножниц?*) и даже в еще большей степени, эти фразы подразумевают, что говорящий сам считает свое

предположение маловероятным, но все же на всякий случай хочет его проверить. Разумеется, эта форма может быть избрана в риторических целях в ситуации, когда говорящий вполне уверен в обстоятельстве, о котором спрашивает:

*Не ты **случайно** мне еще утром обещал помыть посуду?*

Важно отличать вопросительные высказывания с частицами группы *случайно* от высказываний с частицами типа *(а) вдруг*, имеющих гораздо более сложную иллокутивную фунцию. Так, вопрос *Мы **случайно** не опаздываем?* предполагает ответную реплику, содержащую подтверждение или опровержение (*Еще как опаздываем / Вот именно что опаздываем / Надеюсь, нет / Не опаздываем, времени еще полно*). Что же касается высказывания *А **вдруг** мы опоздаем?*, оно также допускает ответы, подтверждающие или не подтверждающие предположение спрашивающего (*Да, похоже, что опоздаем / Да нет, не опоздаем, времени полно*). Однако более типичен ответ совсем в другом ключе: *Ну, тогда поедем на автобусе / Ничего, опоздаем так опоздаем / Это будет катастрофа, ведь эта электричка последняя*. «Гадательное» *вдруг* сближается с такими единицами, как *а что если* и *а ну как*. Последняя единица несколько архаична, но весьма выразительна; она употребляется преимущественно в вопросительных предложениях, но вопросительность и в данном случае не обязательна. Так же ведет себя *(а) (что) если*.

Немецкий язык тоже располагает большим количеством языковых средств, уточняющих иллокутивную установку в речевом акте вопроса. В рассматриваемой функции употребляются многие дискурсивные слова, например: *denn, was, oder, oder was, nicht*. Ср. следующие примеры из параллельного корпуса НКРЯ и немецкоязычного Интернета.

(1) Dann stampfte sie mit dem Fuß auf und rief: „Donnerwetter noch mal! *Hörst du **denn** schwer?*“ [Erich Kästner. Pünktchen und Anton (1931)]
Кнопка топнула ногой и закричала: — *Ты **что**, оглох, окаянный пес?* [Эрих Кестнер. Кнопка и Антон (Е. Вильмонт)]

(2) **а.** *Alles krank **oder was**?* [irre.livejournal.com/161100.html]
— *Все больные **что ли**?*
б. *Sind wir zusammen **oder was**?* [www.welt.de]
*Мы вместе **или как**?*

(3) **а.** «Ах ты мошенник эдакой; ведь я тебе кричал в голос: сворачивай, ворона, направо! *Пьян ты, **что ли**?*» [Н. В. Гоголь. Мертвые души]
„Ach, du Halunke, ich hab dir doch aus aller Kraft zugeschrien: ‚Lenk nach rechts, du Rabenaas!’ *Bist du betrunken, **oder was**?*“ [N. Gogol. Die Toten Seelen]

б. — Что с тобою, мать моя? *С голосу спала, **что ли?..*** [А. С. Пушкин. Пиковая дама]
— Was ist mit dir, du meine Güte? *Hat es dir die Stimme verschlagen, **oder was**?..* [A. Puschkin. Pique Dame]

в. — Что с тобою сделалось, мать моя! *Столбняк на тебя нашёл,* ***что ли****?* Ты меня или не слышишь или не понимаешь?.. [А. С. Пушкин. Пиковая дама]
— Was soll man mit dir machen, du meine Güte? *Träumst du,* ***oder was****?* Entweder hörst du mich nicht, oder du verstehst mich nicht?.. [A. Puschkin. Pique Dame]

(4) «А ты, барин? *Тугиловский,* ***что ли****?*» [А. С. Пушкин. Барышня-крестьянка]
„Und du, Herr? *Bist aus Tugilowo,* ***nicht****?*“ [A. Puschkin. Das Adelsfräulein als Bäuerin]

Надо заметить, что, скажем, фразе *Ты* ***что****, оглох, окаянный пес?* из русского перевода книги Эриха Кестнера «Кнопка и Антон» (пер.: Екатерина Вильмонт) в немецком оригинале соответствует риторический вопрос *Hörst du* ***denn*** *schwer?* При этом в русской и немецкой фразах реализуются разные стратегии говорящего. Если в случае русского *что* говорящий выносит иллокутивную функцию вперед, заявляет ее предварительно при помощи частицы, то в немецкой реплике частица *denn* тяготеет к финалу высказывания.

Обороты *oder was?, nicht?, was?*, а также более позднее *oder?,* характерны для немецкого, однако в русском их следовало бы соотнести в первую очередь с такими оборотами, как *или как? (или что?, или где?* или же *не правда ли?)*

(5) „Eine Geheimkorrespondenz mit Gieshübler“, sagte sie, „Stoff zu neuer Eifersucht für meinen gestrengen Herrn. *Oder nicht*?“ [Theodor Fontane. Effi Briest (1896)]
— Тайная переписка с Гизгюблером, — сказала она. — *Новый повод для ревности, не правда ли,* мой строгий супруг? [Теодор Фонтане. Эффи Брист (Г. Егерман)]

Здесь мы, однако, эти русские единицы рассматривать не будем, так как нас в данном случае в первую очередь интересуют единицы, более близкие к частицам. Что же касается немецкого оборота *oder nicht?,* позже мы вернемся к нему, чтобы отметить одно важное обстоятельство.

Итак, обратимся к немецким частицам. Некоторые из этих дискурсивных слов имеют характер тэгов, превращающих утвердительные высказывания в вопрос. Может показаться, что набор дискурсивных средств в немецком вполне аналогичен тому, что имеется в русском. Между тем, ситуация здесь совершенно другая. Скажем, слово *denn*, которое чаще всего предлагается словарями в качестве перевода русского *разве*, на самом деле имеет гораздо более общее значение. Оно употребляется и в общевопросительных предложениях (*Darfst du denn das?* «Разве тебе это разрешено?»), и в частных вопросах (*Warum denn?* «Почему же?»), и в побудительных высказываниях (*Los denn!* «Ну давай же!») и даже изредка в утверждениях (*Das ist denn auch die Lösung* «Да ведь и это же выход»). Этой своей полифункциональностью *denn* сближается с русской частицей *же*. Несколько огрубляя, можно сказать, что *denn* и *же* — это почти универсальные усилители иллокутивной силы высказывания.

Интересно, что частица *doch*, во многих отношениях сходная с *denn*, в рассматриваемых контекстах употребляется очень ограниченно.[2]

Однако в немецком языке есть и частица, специализирующаяся на общевопросительных предложениях. Это частица *etwa*. При употреблениях в вопросительных предложениях типа *Bist du etwa blöd? = Ты что, дурак?* слово *etwa*, казалось бы, эквивалентно русским вопросительным частицам *что, что же, что ли, разве, неужели, неужто, уж не*. Ср. контексты с (6) по (9) из немецко-русского корпуса параллельных текстов НКРЯ.

(6) Erna sieht mich groß an. „*Soll ich **etwa** allein nachts auf die Straße gehen wie eine Barhure?*“ [Erich Maria Remarque. Der schwarze Obelisk (1956)]
Эрна изумленно смотрит на меня. — ***Что же**, я должна, по-твоему, одна тащиться ночью по улице, как ресторанная шлюха?* [Эрих Мария Ремарк. Чёрный обелиск (В. Станевич)]

(7) „*Halt, um Gottes willen, halt, haben wir uns denn **etwa** wieder übernommen am verdammten Punsch, oder wirkt des Anselmi Wahnsinn auf uns?* Herr Hofrat, was sprechen Sie denn auch wieder für Zeug? [Ernst Theodor Amadeus Hoffmann. Der goldne Topf (1814)]
— *Постойте, ради бога, постойте, напились мы, **что ли**, опять проклятого пуншу или действует на нас сумасшествие Ансельма?* Господин надворный советник, что за чепуху вы опять городите? [Эрнст Теодор Амадей Гофман. Золотой горшок (В. Соловьев)]

(8) „Sie halten Kohler immer noch für schuldig?“ „*Sie **etwa** nicht?*“ [Friedrich Dürrenmatt. Justiz (1985)]
— Вы все еще считаете, что виноват Колер?» — *А вы **разве** нет?* [Фридрих Дюрренматт. Правосудие (В. Герасимов)]

(9) „Erlauben Sie mal, meine Dame, was fällt Ihnen eigentlich ein? *Habe ich es **etwa** nötig, kleine Kinder auszurauben?*“ [Erich Kästner. Emil und die Detektive (1929)]
— Простите, сударыня, но что это вам взбрело в голову? ***Неужели** я похож на человека, который грабит маленьких детей?* [Эрих Кестнер. Эмиль и сыщики (Л. Лунгина)]

Однако все не так просто. В отличие от русских частиц *что, что же, что ли*, немецкое *etwa* может появляться и в несобственно прямой речи (10–11) и даже в гипотаксисе (12). В этих случаях в качестве русских эквивалентов уместны только частицы типа *разве, неужели, неужто, уж не*.

---

[2] Можно сказать *Ihr kommt doch heute Abend?* «Но вы ведь придете сегодня вечером?», но вряд ли *Kommt ihr doch heute Abend?* с вопросительным порядком слов.

(10) Solche Überraschungen hatte ihm sein sonst ganz gefestigter Gesundheitszustand noch nie bereitet. *Wollte **etwa** sein Körper revolutionieren und ihm einen neuen Prozess bereiten, da er den alten so mühelos ertrug?* [Franz Kafka. Der Prozess (1914)]
Никогда его крепкий и в общем здоровый организм не преподносил ему таких сюрпризов. ***Неужто** его тело взбунтовалось, и в нем происходит иной жизненный процесс, не тот, прежний, который протекал с такой легкостью?* [Франц Кафка. Процесс (Р. Райт-Ковалева)]

(11) Ich sehe ihn scharf an. *Sollte er **etwa** selbst ein Auge auf Lisa geworfen haben?* [Erich Maria Remarque. Der schwarze Obelisk (1956)]
Я смотрю на него испытующе. ***Уж не** приглянулась ли она ему самому?* [Эрих Мария Ремарк. Чёрный обелиск (В. Станевич)]

(12) *Doch bereute er, was er durch dieses ewige Suchen und Niemals- oder Immer-Finden, durch dies irdisch-überirdische Fliehen von Begier zu Lust und von Lust zu Begier sonst im Dasein **etwa** versäumt haben mochte?* [Arthur Schnitzler. Casanovas Heimfahrt (1918)]
Но сожалел ли он теперь о том, что упустил в жизни ради этой вечной погони за тем, чего не было нигде и что находилось везде, ради этого земного и неземного метания от желания к наслаждению и от наслаждения к желанию? [Артур Шницлер. Возвращение Казановы (А. Зеленина)]

Эти сочетаемостные особенности немецкой частицы *etwa* связаны со структурой ее многозначности. В своем основном, исходном значение *etwa* — это наречие с семантикой приблизительности (ср. соответствующую словарную статью из НБНС[3]). Идея приблизительности явным образом сохраняется и во всех прочих употреблениях слова *etwa*, в том числе и в рассматриваемых здесь вопросительных предложениях. В этом мы усматриваем базовое отличие семантики *etwa* от ее русских квазиэквивалентов *что, что ли, что же*.

---

[3] **etwa** I *adv* **1.** примерно, около; приблизительно; ~ eine Woche примерно неделю; ~ an dieser Stelle примерно здесь [на этом месте]; so ~ habe ich mir es vorgestellt примерно так я себе это представлял **2.** например; wie ~ как, например; andere Raubtiere wie ~ Wölfe... другие хищники, например, волки... **3.** *террит.* изредка, иногда □ **in ~** примерно, приблизительно; wir stimmen darin in ~ überein в этом наши мнения приблизительно [почти] совпадают II *prtc mod* **1.** (*в вопросах с отрицанием, не несёт фразового ударения*) разве; что ли (*разг.*); kennen Sie ihn ~ nicht? разве вы его не знаете?, *при ударении на глаголе* вы его не знаете разве [что ли]? **2.** (*в вопросах без отрицания, не несёт фразового ударения*) может быть, разве; bist du ~ krank? может быть, ты болен?, ты не болен?; ist das ~ wichtig? разве это важно? **3.** (*в условных предложениях*) в случае; wenn er ~ doch kommen sollte если он вдруг придёт □ **nicht ~** совсем не, вовсе не; das waren nicht ~ Beweise, sondern bedeutungslose Daten это были совсем [никакие] не доказательства, а ничего не значащие данные; er wollte das Rad nicht ~ stehlen, sondern nur ausleihen он вовсе не хотел красть велосипед, он хотел просто взять его на прокат; ich bin nicht ~ dagegen, aber... я совсем не против, но...; (**doch) nicht ~** *в вопросах с прямым порядком слов* уж не... ли; du bist doch nicht ~ Pfarrer? уж не священник ли ты?

Заметим, что *etwa* может соответствовать не только частицам типа *разве*, но и частицам типа *(а) вдруг*. Ср.: *А вдруг он опоздает?* Подобно *вдруг*, *etwa* может встречаться и в условных контекстах: *wenn er etwa doch kommen sollte* «если он вдруг придёт».

Понятно, что идея приблизительности не препятствует употреблениям частицы *etwa* в свободном косвенном дискурсе и придаточных предложениях, передающих прямую речь персонажа, поэтому эта частица не так жестко, как ее русские аналоги, привязана к диалогическому режиму.

Теперь вернемся к оборотам *oder was?* и *oder nicht?* и отметим их важное отличие от *etwa*. Оборотам *oder was?* и *oder nicht?* предшествует высказывание, основанное на изначальных установках. А вопросы с *etwa*, как и с *что* — это предположения основанные на новых, только что замеченных признаках ситуации, на только что полученной информации, которая заставляет пересмотреть изначальные установки.

Во фразах *Du kommst mit, oder nicht? = Ну что, ты идешь? — Kommst du etwa mit? = Ты что тоже с нами идешь?* реализуются совершенно разные коммуникативные ситуации. В первом случае коммуниканты договаривались идти вместе, а адресат не готов. Во втором они ни о чем не договаривались, а адресат ведет себя так, что можно подумать, что он тоже собирается идти.

Bleibst du etwa zu Hause? = Ты что дома остаешься? синонимично *Du kommst mit, oder nicht?*, только более возмущенное. Фокусируется альтернативная ситуация. Таким образом, оказывается, что различие между немецкими дискурсивными оборотами аналогично тому противопоставлению, которое имеется и между русскими дискурсивными средствами.

***

Материал частиц позволяет ясно увидеть разницу между переводимостью и семантическим тождеством. При богатом репертуаре вопросительных частиц и в русском, и в немецком языках для каждого высказывания можно подобрать хороший эквивалент, причем с частицей. Однако из этого не следует, что сами частицы семантически эквивалентны. Анализ показывает, что смысловое сходство между самими частицами довольно отдаленное. Это лишний раз подтверждает мысль, высказанную еще Якобсоном — переводимы высказывания, но не слова: «Слово или фразеологический оборот (иначе говоря: единицу кода более высокого уровня) можно полностью интерпретировать только через эквивалентную комбинацию кодовых единиц, то есть через сообщение, относящееся к этой единице. <…> Точно так же на уровне межъязыкового перевода обычно нет полной эквивалентности между единицами кода, но сообщения, в которых они используются, могут служить адекватными интерпретациями иностранных кодовых единиц или целых сообщений» [Якобсон 1978: 16].

В нашем случае имеет место именно такая ситуация: сравниваемые частицы во многих случаях функционально эквивалентны, но семантически они различны. *Etwa* — в силу связи с идеей приблизительности — скорее смягчает

иллокутивную установку, а русские *что, что ли, как* содержат прямую апелляцию к собеседнику с предварительно заявленной вопросительностью и тем самым увеличивают иллокутивную силу. Однако и то и другое позволяет как-то «оживить» вопрос, продемонстрировать включенность говорящего в ситуацию, и именно этим определяется функциональная близость.

Литература

1. *Булыгина Т. В., Шмелев А. Д.* О семантике частиц разве и неужели // НТИ. Сер. 2. 1987. № 10. С. 23–30.
2. *Булыгина Т. В., Шмелев А. Д.* Языковая концептуализация мира (на материале русской грамматики). М.: Языки русской культуры, 1997.
3. *Добровольский Д. О., Левонтина И. Б.* Диалогические частицы: русско-немецкие соответствия // Логический анализ языка. Моно-, диа-, полилог в разных языках и культурах / Отв. ред. член-корр. РАН Н. Д. Арутюнова. М.: Индрик, 2010. С. 93–103.
4. *Добровольский Д. О., Левонтина И. Б.* О синонимии фокусирующих частиц (на материале немецкого и русского языков) // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2012». Выпуск 11 (18): В 2 т. Том 1. М.: Изд-во РГГУ, 2012. С. 138–149.
5. *Левонтина И. Б.* Дискурсивные слова в вопросительных предложениях // Die Welt der Slaven (в печати).
6. НБНС — Новый большой немецко-русский словарь. В 3 т.: около 500 000 лексических единиц / под общим руководством Д. О. Добровольского. М.: Астрель, 2008–2010.
7. *Рестан П.* Синтаксис вопросительного предложения. Осло, Берген, Тромсё 1972.
8. *Якобсон Р.* О лингвистических аспектах перевода // Вопросы теории перевода в зарубежной лингвистике. М.: Международные отношения, 1978. С. 16–24.

References

1. *Bulygina T. V., Shmelev A. D.* (1987), On semantics of the particles *razve* and *neuzheli* [O semantike chastits *razve* i *neuzheli*], Scientific and Technical Information Processing, 2 [Nauchno-tehnicheskaja informatsija, serija 2], No. 10, pp. 23–30.
2. *Bulygina T. V., Shmelev A. D.* (1997), Language conceptualization of the world (Russian grammar) [Jazykovaja kontseptualizatsija mira (na materiale russkoj grammatiki)], Jazyki russkoj kul'tury, Moscow.
3. *Dobrovol'skij D. O., Levontina I. B.* (2010), Dialogue particles: Russian-German correspondences [Dialogicheskie chastitsy: russko-nemetskie sootvetstvija], Logical analysis of language [Logicheskij analiz jazyka], Indrik, Moscow, pp. 93–103.

4. *Dobrovol'skij D. O., Levontina I. B.* (2012), Synonymous focus particles in German and Russian [O sinonimii fokusirujushchih chastits], Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue 2012", Issue 11 (18) [Komp'juternaja Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2006"], Vol. 1, Bekasovo, pp. 138–149.
5. *Levontina I. B.* (in print), Discourse words in question sentences [Diskursivnye slova v voprositel'nyh predlozhenijah], Die Welt der Slaven.
6. *NGRCD* (2008–2010) — New German-Russian Comprehensive Dictionary [Novyj bol'shoj nemetsko-russkij slovar'], Astrel', Moscow.
7. *Restan P.* (1972), The syntax of the question sentence [Sintaksis voprositel'nogo predlozhenija], Oslo, Bergen, Tromsö.
8. *Jakobson R.* (1978), On linguistic aspects of translation [O lingvisticheskih aspektah perevoda], Translation theory in linguistics abroad [Voprosy teorii perevoda v zarubezhnoj lingvistike], Mezhdunarodnye otnoshenija, Moscow, pp. 16–24.

МОДАЛЬНЫЕ ПРЕДИКАТЫ И СОСЛАГАТЕЛЬНОЕ НАКЛОНЕНИЕ[1]

Добрушина Н. Р. (nina.dobrushina@gmail.com)

Национальный исследовательский университет
Высшая школа экономики, Москва, Россия

В статье рассматриваются сочетания модального глагола *мочь* и прилагательного *должен* с частицей *бы*. Показано, что эти модальные предикаты в форме сослагательного наклонения обнаруживают поведение, отличное от обычных глаголов: сослагательное наклонение нередко синонимично индикативу, и в контекстах, где для других глаголов сослагательное наклонение является обязательным, предикаты *мочь* и *быть должным* стоят в форме индикатива. Основным фактором опущения частицы сослагательного наклонения является эпистемическое значение этих модальных предикатов. В статье используются данные Национального корпуса русского языка.

Ключевые слова: модальность, наклонение, сослагательное наклонение, условные конструкции

MODALS AND THE SUBJUNCTIVE

Dobrushina N. R. (nina.dobrushina@gmail.com)

National Research University Higher School of Economics,
Moscow, Russia

I consider constructions that involve the modal verb *moch'* or the modal adjective *dolzhen* and the subjunctive particle *by*. I argue that, with respect to the subjunctive, these modals behave differently from regular verbs. Their subjunctive is often functionally identical to the indicative; in contexts where other verbs obligatorily take the subjunctive form, these two predicates may use the indicative. The main factor that controls omissibility of the subjunctive particle is shown to be an epistemic interpretation. I consider some typical cases where the subjunctive and the indicative are synonymous

[1] Работа выполнена в рамках проекта корпусного описания русской грамматики «Русграм» (http://rusgram.ru), поддержанного РГНФ (14-04-00264, рук. Плунгян В. А. Семантико-синтаксический компонент интегрированного корпусного описания русской грамматики) и Программой фундаментальных исследований Президиума РАН «Корпусная лингвистика», 2012–2014.

for these predicates, and those where they are not. Thus, in the apodosis of conditional constructions the particle is often omitted, although, in general, Russian prefers a symmetrical use of the subjunctive in both protasis and apodosis. On the other hand, when in the protasis, the particle is not omitted. The subjunctive is often used with the modals for pragmatic purposes, such as politeness. The paper is based on the data from the Russian National Corpus.

Key words: modality, mood, subjunctive, conditional clauses

Введение

В ряде европейских языков наблюдается такое явление, как употребление модальных глаголов в ирреальном наклонении с некоторыми нестандартными семантическими эффектами (Holvoet 2010: 429–430, Palmer 2001: 2011–2014, Bybee 1995: 506). Аксель Хольвут, обсуждая ирреальное маркирование модальных глаголов в латышском и в литовском языках, высказывает мнение о том, что наличие маркеров ирреальности на модальных глаголах в балтийских языках поддерживается влиянием немецкого языка, а отсутствие такового — влиянием русского языка (Holvoet 2010: 430), таким образом предполагая, что для русского языка ирреальное маркирование модальных предикатов не характерно.

Между тем русские модальные предикаты, особенно глагол *мочь*, нередко сочетаются с частицей сослагательного наклонения. При этом наблюдаются некоторые интересные семантические особенности:

- форма сослагательного наклонения глагола *мочь* и прилагательного *должен* может быть заменена на форму индикатива без всякого изменения в значении:

(1) *Он **мог бы** вернуться, но не захотел = Он **мог** вернуться, но не захотел*

- глагол *мочь* и прилагательное *должен* могут не иметь маркирования частицей сослагательного наклонения в таких контекстах, где другие глаголы должны его иметь:

(2) *Если бы он захотел, он **мог бы*** [мог] *вернуться*

(3) *Если бы он захотел, он **постарался бы*** [*постарался] *вернуться*.

Эта статья посвящена исследованию нетривиальных семантических эффектов, которые обнаруживаются при взаимодействии модальных предикатов *мочь* и *должен* с граммемой сослагательного наклонения.

1. Когда сослагательное наклонение и индикатив синонимичны

1.1. Независимые предикации

Особенностью формы сослагательного наклонения предикативов *мочь* и *должен* является то, что в большинстве контекстов она взаимозаменима с формой индикатива. В следующих примерах ситуация является контрфактивной, то есть такой, которая никогда не имела места в действительности и никогда не будет иметь.

(4) *О, мои неслучившиеся дети: один* ***мог бы*** [ОК: мог] *стать китайским божком, другой маркизиком.* [Кира Сурикова. ДТП (2003)]

(5) ***Мог бы*** [ОК: мог]***,*** *конечно, разбудить всех и занять свое место, но решил не связываться.* [Михаил Гиголашвили. Чертово колесо (2007)]

(6) *А ведь эти люди* ***должны были бы*** [ОК: должны были] *пойти в малый бизнес.* [Модест Колеров, Евгений Гонтмахер. «Государство не должно светиться» // «Отечественные записки», 2003]

В таких случаях обычные глаголы могут иметь только форму сослагательного наклонения. Использование индикатива изменит смысл предложения:

(7) *О, мои неслучившиеся дети: один* ***стал бы*** [*стал] *китайским божком, другой маркизиком.*

(8) ***Разбудил бы*** [*разбудил]***,*** *конечно, всех и* ***занял бы*** [*занял] *свое место, но решил не связываться.*

(9) *А ведь эти люди* ***пошли бы*** [*пошли] *в малый бизнес.*

Существуют также контексты, где сослагательное наклонение не может быть заменено на прошедшее время индикатива, но допускает замену на настоящее время. Это происходит в тех примерах, где ситуация имеет референцию к настоящему или будущему и тем самым является гипотетической, то есть имеет шансы на реализацию:

(10) *Дмитриев подумал, что* ***мог бы*** [*мог / ОК: может] *завтра переселиться в эту трехкомнатную квартиру, видеть по утрам и по вечерам реку, село, дышать полем, ездить на работу автобусом до Серпуховки, оттуда на метро, не так уж долго.* [Юрий Трифонов. Обмен (1969)]

(11) *а. Он* ***мог б****ы* [мог] *передвигаться самостоятельно, но у него не было денег на протезы.*

б. Он ***мог бы*** [*мог / ОК: может] *передвигаться самостоятельно, но у него нет денег на протезы.*

Некоторое семантическое различие между модальными предикатами *мочь* и *должен* в индикативе и в сослагательном наклонении в гипотетическом значении есть, но весьма слабое. Обе формы обозначают ситуацию, которая имеет шансы на реализацию, но сослагательное наклонение используется для ситуаций, реализация которых является лишь одной из многих возможностей, в то время как индикатив используется для ситуаций, реализация которых — наиболее возможный вариант. Разница, таким образом, в степени вероятности осуществления ситуации: более низкая — при сослагательном наклонении, более высокая — при индикативе (см. Hansen 2005: 228 о том, что *должен был бы* выражает слабую необходимость по сравнению с *должен*, обозначающим высокую степень).

(12) *Другого, более надежного способа удержать Чечню от послевыборного хаоса, обид и раздоров сегодня нет. Кроме того, работа с претендентами* ***могла бы*** *дать ценный материал для точного понимания ситуации, заложить основы строительства мирного гражданского общества, стать своего рода кадровой революцией.* [Аднан Музыкаев. Выборы президента Чечни — игра с огнем? // «Аргументы и факты», 2003]

(13) *«Евросеть» подала заявку на размещение глобальных депозитарных расписок на Лондонской бирже. Предприниматель Александр Мамут* ***может продать*** *принадлежащие ему 50 % плюс одна акция либо меньший пакет в зависимости от спроса на акции компании. Кроме того, в ходе размещения ритейлер планирует допэмиссию в размере около 140 млн долл., необходимых на развитие компании.* [Антон Бурсак. «Евросеть» подала заявку на размещение акций на Лондонской бирже (2011.03.23) // http://www.rbcdaily.ru/2011/03/23/media/562949979916514.shtml, 2011]

(14) *То есть, правильным решением было бы сосуществование двух линий — внедряемая в производство консервативная и развивающаяся нормальными темпами и без лишней спешки революционная. Вторая* ***должна бы*** *лет через 20–30–40 заменить в случае удачи первую. Если же от консервативных технологий отказаться напрочь, то что мы скажем людям?* [коллективный. Российский атом. Проект ПРОРЫВ (2012)]

(15) *Не случайно говорилось, что в литературу он вошел раньше, чем в Союз писателей. Впрочем, так и* ***должно бы быть****. Если б так было!* [Г. Я. Бакланов. Разное // «Знамя», 2002]

Интересно сочетание сослагательного наклонения с наречием *непременно* — в следующем примере обозначена высокая степень необходимости при низкой вероятности осуществления события:

(16) *А уж здесь-то конкурс* ***непременно должен бы*** *состояться.* [Давид Константиновский. Свет мой, зеркальце! Скажи… // «Отечественные записки», 2003]

Глагол *мочь* в сослагательном наклонении часто используется прагматически, для смягчения категоричности высказывания — в рамках стратегии отрицательной вежливости:

(17) *Ты* ***не могла бы*** *дать мне рецепт этого пирога?* [Коллекция анекдотов: семья (1970–2000)]

(18) *Вы* ***могли бы*** *перед собранием повторить все то, что рассказали нам?* [Михаил Елизаров. Библиотекарь (2007)]

Здесь тоже невозможна замена на индикатив прошедшего времени (поскольку ситуация является гипотетической, а не контрфактивной), но возможна — на индикатив настоящего:

(19) *Ты* ***не могла бы*** [*не могла / ОК: не можешь] *дать мне рецепт этого пирога?*

(20) *Вы* ***могли бы*** [*могли / ОК: можете] *перед собранием повторить все то, что рассказали нам?*

Разница между сослагательным наклонением и индикативом здесь трансформируется из степени вероятности в степень вежливости — при более дистантных отношениях предпочтительно сослагательное наклонение. Так, в типичных конструкциях для выражения просьбы вида «ты не мог бы…?» или «ты не можешь…? » сослагательное наклонение заметно частотнее, если подлежащим является местоимение «вы» (см. таблицу 1), поскольку оно обычно обозначает вежливое обращение. Запрос формулировался следующим образом: «ты | вы не мочь расстояние 1–8-и?», где перед знаком вопроса может быть любое слово под отрицанием (в данном запросе было слово *и*), в выдаче вручную выбирались примеры со значением просьбы. Возможно, что в некоторых случаях «вы» означает обращение к нескольким людям, с которыми говорящий на «ты», но это не всегда удается проверить.

Таблица 1. Распределение наклонений при выражении просьбы

	ты	**вы**
индикатив	43% (41)	21% (28)
сослагательное наклонение	57% (54)	79% (108)
всего примеров	*95*	*136*

1.2. Сослагательное наклонение модальных предикатов в аподозисе условных конструкций

Важным доказательством того, что модальные предикаты *мочь* и *должен* отчасти синонимичны сослагательному наклонению, является распространенный тип условных конструкций, в котором протазис маркирован сослагательным наклонением, а в аподозисе находится глагол *мочь* (чаще) или прилагательное *должен* (реже) без частицы сослагательного наклонения:

(21) ***Если бы*** *в мои планы входило украсть что-нибудь, я* ***мог*** *вынести даже прилавок.* [Андрей Геласимов. Год обмана (2003)]

(22) *И* ***если бы*** *не взяли,* ***если бы*** *мы продолжали основными силами штурмовать ВОВД, русские* ***могли*** *накрыть этот квадрат, уничтожить и наших, и своих.* [Герман Садулаев. Шалинский рейд (2009) // «Знамя», 2010]

(23) *И Дарья радовалась: конечно, бабушка и без нее бы обошлась, но если бы она не взялась помогать, в семье* ***могла*** *произойти ссора.* [Наталия Ермильченко. Генеральная уборка // «Мурзилка», 2002]

(24) *Потому что если бы они померли сами, то* ***должны были*** *где-нибудь валяться.* [Андрей Геласимов. Степные боги (2008)]

(25) *— Если бы такие проверки проходили, то информация* ***должна была быть*** *предоставлена Совету Думы, — заметила Слиска.* [«Известия», 2002.02.05]

При этом в целом для русских УК с сослагательным наклонением характерно маркирование наклонением обеих частей. Действительно, случайная выборка в подкорпусе с 1970 года показывает, что из 270 условных конструкций, содержащих в протазисе *если бы* и в аподозисе глагол *мочь*, в 63 отсутствует частица *бы,* то есть в 23 % случаев *бы* опускается. В то же время в выборке из 300 условных конструкций, содержащих в протазисе *если бы* и не ограниченных типом предиката аподозиса, частица *бы* отсутствует лишь в 14, и в половине из них имеются модальные слова *мочь, можно* или *должен*:

(26) *Если бы она сняла выигрыш, то на сумку потом* ***можно было*** *у кого-нибудь перехватить.* [Андрей Геласимов. Дом на Озерной (2009)]

(27) *На какую высоту над поверхностью воды* ***должен был*** *выскочить мячик, если бы сопротивление воды* (и воздуха) *отсутствовало?* [Владимир Лукашик, Елена Иванова. Сборник задач по физике. 7–9 кл. (2003)]

(28) *Если бы боевики орудовали плоскогубцами, то* ***выдирали*** *те зубы, к которым легче доступ, то есть передние.* [Токарева Виктория. Своя правда // «Новый Мир», 2002]

Таблица 2. Опущение частицы ***бы*** в аподозисе УК, протазис которых содержит сослагательное наклонение

	частица *бы* в аподозисе есть	**частицы *бы* в аподозисе нет**
УК с глаголом *мочь* в аподозисе	77% (207)	23% (63)
УК с произвольным предикатом в аподозисе	93% (286)	7% (14)
всего	*300*	*300*

Заметим, что опущение частицы *бы* в аподозисе УК при глаголе *мочь* еще более вероятно в том случае, если протазис выражен не сослагательным наклонением, а императивом. Так, если конструкция (а) оценивается носителями как допустимая, но разговорная, то конструкция (б) принимается без всяких оговорок:

(29) *а. Петр Ильич говорил, что* ***если бы*** *Столыпин остался жив, он мог спасти Россию.*
б. Петр Ильич говорил, что Столыпин, ***останься*** *он жив, мог спасти Россию.*

Видимо, это связано с тем, что в русском языке существует довольно сильная тенденция к согласованию наклонений в обеих частях УК, меж тем как для конструкций с императивом в протазисе согласование наклонений невозможно в принципе, и потому индикатив возможен в той же степени, что и сослагательное наклонение.

2. Когда индикатив и сослагательное наклонение несинонимичны

Итак, сослагательное наклонение модальных предикатов *мочь* и *должен* часто может быть заменено индикативом без изменения значения. Однако существует несколько условий, при которых замена невозможна.

Во-первых, индикатив несинономичен сослагательному наклонению, в случае если глагол *мочь* употреблен в значении «быть способным, быть в состоянии, обладать умением»:

(30) *Если бы ему вовремя сделали операцию, он* ***бы мог*** [*мог] *передвигаться самостоятельно.*

Сравним этот пример с рассмотренными выше примерами, где *мочь* имеет эпистемическое значение — «иметь шансы случиться»:

(31) *Если бы она не помогала бабушке, в семье* ***могла бы*** [могла] *произойти ссора.*

Таким образом, при эпистемическом значении индикатив и сослагательное наклонение синонимичны, при других значениях — нет. Нужно отметить, что в составе условных конструкций значительно чаще встречается эпистемическое значение, чем значение «быть способным, быть в состоянии, обладать умением». В подавляющем большинстве случайно выбранных примеров условных конструкций с глаголом *мочь* в аподозисе замена одного наклонения другим возможна.

Во-вторых, невозможность замены может быть связана с итеративным значением глагола *мочь*, когда он обозначает ситуацию, которая время от времени имела место:

(32) *Правда, изредка и он **мог** /* [* мог бы] *раздражиться, стать на короткое время резким и даже неприятным — когда слишком уж доводили.* [Константин Ваншенкин. Писательский клуб (1998)]

Сравним этот пример со следующим, где замена возможна:

(33) *Я стоял ниже и **мог** /* [ОК: мог бы] *легко использовать это преимущество: сделать ложную уступку и тут же бросить его через бедро.* [Константин Ваншенкин. Писательский клуб (1998)]

Чем отличаются эти два примера? В первом, где замена невозможна, конструкция с глаголом *мочь* обозначает ситуацию, которая периодически имела место в реальности (ср. Падучева 2014: «Показатель внешней возможности может выражать экзистенциальную квантификацию»):

(34) *У него был трудный характер. Он **мог** сделать резкое замечание, отругать, накричать.* = [Бывало такое, что он делал резкое замечание, ругал, кричал]

Во примере (33) глагол *мочь* в прошедшем времени обозначает потенциальную ситуацию, которая не имела места в действительности, а относится лишь к воображаемому миру. В этом случае индикатив заменим на сослагательное наклонение:

(35) *Я **мог** легко использовать это преимущество. = Я **мог бы** легко использовать это преимущество.*

(36) *В тот момент он **мог** сделать резкое замечание, накричать, но удержался. = В тот момент он **мог бы** сделать резкое замечание, накричать, но удержался.* [В реальной ситуации он не кричал, в альтернативном мире такое можно вообразить]

В-третьих, глагол *мочь* с отрицанием реже допускает замену индикатива сослагательным наклонением. Так, невозможна замена в том случае, если ситуация имеет конкретную референцию, то есть в некоторый определенный момент прошлого участник не был в состоянии осуществить действие:

(37) *Кенни тут же воспользовался открывшимся коридором и проскользнул вперед, а обескураженный бразилец еще долго* ***не мог обогнать*** [**не мог бы*] *Lola Паписа и в конце концов откатился на итоговое седьмое место.* [Дмитрий Ситник. CART: второй интернационал (2001) // «Формула», 2001.09.1]

Сослагательное наклонение при отрицательном глагол *мочь* уместно, если ситуация не является конкретно-референтной, то есть нельзя определить точный момент в прошлом, когда она имела место:

(38) *Теперь мальчишка уже не* ***мог*** [ОК: ***не мог бы***] *выскочить на улицу голым в январские морозы, а вынужден был надевать брюки, обувь, сюртучок.* [Александр Волков. Человек между лесом и волком // «Знание — сила», 2003]

Модальное прилагательное *должен* реже допускает замену индикатива на сослагательное наклонение. Это связано с тем, что, как было сказано выше, граммема сослагательного наклонения ослабляет степень вероятности того, что событие осуществится. Между тем должен, в отличие от мочь, обозначает высокую степень вероятности; поэтому в сочетании с сослагательным наклонением оно обычно почти однозначно подразумевает, что ситуация имеет мало шансов быть реализованной или вообще контрфактивна. Поэтому сослагательное наклонение совершенно уместно в примере (а), а в примере (б) создает смысл, которого исходное предложение не имеет (в):

(39) *На вышках* ***должны были*** [=должны были бы] *дежурить часовые, но их никогда там не было.* [Р. Б. Ахмедов. Промельки (2011) // «Бельские Просторы»]

(40) *К Чернышеву* ***должна была*** *приехать невеста Наташа, молоденькая девушка с короткой стрижкой, он всем показывал ее фотографию и хвастал, что она дочь известного писателя.* [Р. Б. Ахмедов. Промельки (2011) // «Бельские Просторы»]

(41) *К Чернышеву* ***должна была бы*** *приехать невеста Наташа = К Чернышеву должна была приехать невеста Наташа, но не приехала.*

2.1. Сослагательное наклонение модальных предикатов в некоторых подчиненных предикациях

Интересно, что когда глагол *мочь* употребляется в протазисе условной конструкции, где аподозис выражен сослагательным наклонением, то описанное выше явление не наблюдается: сослагательное наклонение не может быть заменено индикативом:

(42) *Если* ***бы она могла*** [*если она могла***]***, то умерла бы скорее.* [Роман Сенчин. Елтышевы (2008) // «Дружба Народов», 2009]

(43) *Если* ***бы мы могли [****если мы могли**] наполнить им чайную ложечку и взвесить её, то на Земле она весила бы около тонны!* [Е. Левитан. Космические ужастики // «Наука и жизнь», 2008]

(44) — *Если* ***бы дядя Максим мог [****если дядя Максим мог**] высказать свое мнение, он наверняка предпочел бы вас.* [Михаил Елизаров. Библиотекарь (2007)]

(45) — ***Если бы ты только могла [****если ты только могла**] себе представить, как я тебя ненавижу, — печально признался он.* [Татьяна Тронина. Никогда не говори «навсегда» (2004)]

Представляется, что применение этого правила не зависит от значения глагола *мочь*. Сравним следующие примеры, где глагол *мочь* имеет эпистемическое употребление, обозначая степень вероятности ситуации:

(46) *Если бы Ноздрев не пил и не буянил, он* ***мог бы [мог]*** *стать чемпионом по шашкам и шахматам* [имел шансы стать чемпионом]. [С. Атасов. 1000 золотых анекдотов (2003)]

(47) ***Если бы Ноздрев мог*** [*если Ноздрев мог] *стать чемпионом по шашкам, он бы не пил и не буянил.*

По-видимому, запрет на индикатив в протазисе таких условных конструкций является синтаксическим правилом.

В определительных придаточных предложениях гипотетического типа, где использование сослагательного наклонения регулируется рядом факторов, наблюдающихся в главном предложении (наличие термового отрицания, семантический тип предиката, нереферентность имени и другие — Dobrushina 2010), модальные предикаты *мочь* и *должен* позволяют замену индикативом, причем если предикат имеет чисто эпистемическое значение, то он синонимичен основному глаголу в сослагательном наклонении. Можно сформулировать более сильное утверждение: названные выше условия требуют, чтобы в определительном придаточном было либо сослагательное наклонение, либо модальный предикат:

(48) *Честно скажу вам, Елизавета Григорьевна, я не вижу причин, которые* ***могли бы*** [ОК: могут] *заставить* [ОК: заставили бы] *вашего шефа отказаться от этого плана.* [Зиновий Юрьев. Смертельное бессмертие // «Наука и жизнь», 2007]

(49) *Нет такого предмета, который я* ***не могла бы*** [ОК: не могу] *сделать сама и за меня его* ***должен был бы*** [ОК: должен был] *делать кто-нибудь другой.* [Наталья Корнеева. Оксана Ярмольник: «С удовольствием построила бы еще один дом» (2004) // «Homes & Gardens», 2004.03.02]

3. Заключение

Итак, были рассмотрены несколько случаев, когда при использовании с модальными предикатами *мочь* и *быть должным* нейтрализуется значение сослагательного наклонения: оно оказывается синонимично индикативу и может быть заменено на него. Эта особенность модальных предикатов связана с их значением, которое подразумевает, что ситуация не является вполне реализованной: “Modal verbs, whether they express desire, obligation, intention or ability, have in common the semantic property that they do not imply the completion of the action or event expressed by the infinitive with which they occur” (Bybee 1995: 505). В терминологии Тальми Гивона, эти предикаты обладают встроенной ирреальностью (inherent irreality — Givon 1984: 306). Однако этого объяснения недостаточно, поскольку разные типы употребления модальных предикатов по-разному взаимодействуют с сослагательным наклонением.

Хорошо известно, что модальные предикаты *мочь* и *должен* могут иметь эпистемическое значение (вероятностное, предположительное — Апресян 1995: 191–193) и другое, которое в терминологии ван дер Ауверы и Плунгяна называется неэпистемическим (van der Auwera, Plungian 1998: 81–82), а в терминологии Ю. Д. Апресяна — собственно модальным. Предварительное исследование показало, что синонимия сослагательного наклонения с индикативом в первую очередь характерна для эпистемических контекстов. Однако неэпистемические контексты в разной степени допускают опущение частицы сослагательного наклонения. Так, в примере (50) представлено эпистемическое значение (был шанс, что в семье произойдет ссора), в примерах (51) и (52) — неэпистемическое (был в состоянии бросать через бедро; был в состоянии передвигаться), однако в примере в) индикатив исключен, а в примере б) допустим.

(50) *Если бы она не помогала бабушке, в семье* ***могла бы*** [ОК: могла] *произойти ссора.*

(51) *Если бы я стоял ниже, я* ***мог бы*** [?ОК: мог] *бросать его через бедро.*

(52) *Если бы ему вовремя сделали операцию, он* ***мог бы*** [*мог] *передвигаться самостоятельно.*

Взаимодействие модальных предикатов с сослагательным наклонением, таким образом, требует дальнейших исследований.

Литература

1. *Апресян Ю. Д.* (1995). Избранные труды. Т. 2. Интегральное описание языка и системная лексикография. М.: Школа «Языки русской культуры».
2. *Падучева Е. В.* (2014). Модальность. Материалы для проекта корпусного описания русской грамматики (http://rusgram.ru). На правах рукописи.
3. *Bybee, J.* (1995). The semantic development of past tense modals in English. Bybee, Joan; Fleischman, Suzanne. Modality in Grammar and Discourse. Philadelphia: John Benjamins, 503–517.
4. *Dobrushina, Nina.* Subjunctive in relative clauses. (2010). Oslo studies in language 2: 181–210.
5. *Holvoet, A.* (2010). Mood in Latvian and Lithuanian. Mood in the Languages of Europe (Vol. 120). John Benjamins Publishing.
6. *Givón, Talmy.* (1984). Syntax Vol. I. The Hague: John Benjamins Publishing Company.
7. *Palmer F. R.* 2001. Mood and Modality. Cambridge: Cambridge University Press.
8. *Van der Auwera, J., & Plungian, V. A.* (1998). Modality's semantic map. Linguistic typology, 2(1), 79–124.

References

1. *Apresjan, Ju. D.* (1995). Izbrannye trudy. T. 2. Integral'noe opisanie jazyka i sistemnaja leksikografija. [Selected Works, Volume II. Integral Description of Language and System Lexicography]. Jazyki russkoj kul'tury, Moskva.
2. *Paducheva E. V.* (2014). Modal'nost'. Materialy dlja proekta korpusnogo opisanija russkoj grammatiki (http://rusgram.ru). Manuscript.
3. *Bybee, J.* (1995). The semantic development of past tense modals in English. Bybee, Joan; Fleischman, Suzanne. Modality in Grammar and Discourse. Philadelphia: John Benjamins, 503–517.
4. *Dobrushina, Nina.* Subjunctive in relative clauses. (2010). Oslo studies in language 2: 181–210.
5. *Holvoet, A.* (2010). Mood in Latvian and Lithuanian. Mood in the Languages of Europe (Vol. 120). John Benjamins Publishing.
6. *Givón, Talmy.* (1984). Syntax Vol. I. The Hague: John Benjamins Publishing Company.
7. *Palmer F. R.* 2001. Mood and Modality. Cambridge: Cambridge University Press.
8. *Van der Auwera, J., & Plungian, V. A.* (1998). Modality's semantic map. Linguistic typology, 2(1), 79–124.

РАСШИРЕНИЕ ЗАПРОСА В ИНФОРМАЦИОННОМ ПОИСКЕ: ЧТО МЫ МОЖЕМ УЗНАТЬ ИЗ ГЛУБИННОГО АНАЛИЗА ЗАПРОСА?

Ермакова Л. М. (liana.ermakova@irit.fr)

Institut de Recherche en Informatique de Toulouse, CNRS UMR 5505, Université de Toulouse, Тулуза, Франция;
Пермский государственный национальный исследовательский университет, Пермь, Россия

Мот Ж. (josiane.mothe@irit.fr)

Institut de Recherche en Informatique de Toulouse, CNRS UMR 5505, Université de Toulouse, Тулуза, Франция

Овчинникова И. Г. (ira.ovchi@gmail.com)

Пермский государственный национальный исследовательский университет, Пермь, Россия

Одна из основных задач информационного поиска—извлечение документов, релевантных информационной потребности пользователя, выраженной запросом. Зачастую пользовательские запросы не превосходят 3 слов, что усложняет задачу. Многочисленные исследования показали, что автоматическое расширение запроса в среднем повышает точность, несмотря на то, что для некоторых запросов результаты ухудшаются. В статье предлагается новый метод автоматического расширения запроса, основанный на оценки важности слов-кандидатов, определяемой силой их связи со словами из запроса. Предлагаемый метод комбинирует локальный анализ, а именно обратную связь по релевантности, и глобальный анализ коллекции документов. Оценка метода была произведена на международных тестовых коллекциях, согласно установленным метрикам. Полученные результаты были сравнены с одной из лучших моделей, описанных в литературе. Системы показали сравнимые результаты в среднем. Однако глубинный анализ исходных и расширенных запросов позволил сделать выводы, которые могут помочь в исследовании данной области.

Ключевые слова: информационный поиск, расширение запроса, анализ слов запроса, обратная связь по релевантности, глобальный анализ, совстречаемость

QUERY EXPANSION IN INFORMATION RETRIEVAL: WHAT CAN WE LEARN FROM A DEEP ANALYSIS OF QUERIES?

Ermakova L. M. (liana.ermakova@irit.fr)
Institut de Recherche en Informatique de Toulouse, CNRS UMR 5505, Université de Toulouse, France;
State National Research University, Perm, Russia

Mothe J. (josiane.mothe@irit.fr)
Institut de Recherche en Informatique de Toulouse, CNRS UMR 5505, Université de Toulouse, France

Ovchinnikova I. G. (ira.ovchi@gmail.com)
Perm State National Research University, Perm, Russia

Information retrieval aims at retrieving relevant documents answering a user's need expressed through a query. Users' queries are generally less than 3 words which make a correct answer really difficult. Automatic query expansion (QE) improves the precision on average even if it can decrease the results for some queries. We propose a new automatic QE method that estimates the importance of expansion candidate terms by the strength of their relation to the query terms. The method combines local analysis and global analysis of texts. We evaluate the method using international benchmark collections and measures. We found comparable results on average compared to the Bo2 method. However, we show that a deep analysis of initial and expanded queries brings interesting insights that could help future research in the domain.

Key words: information retrieval, query expansion, analysis of query terms, relevance feedback, global analysis, co-occurrence

1. Introduction

Information Retrieval (IR) systems aim at retrieving information that answers a user's needs/he expresses through a query. Retrieving relevant information to a query implies a two-step process: off line, the system indexes documents, generally using a bag of words representation; online, the system computes the similarity between the user's query and the document representations (indexing terms) to retrieve the most similar documents.

IR systems have to face the problem of query term ambiguity inherent in natural language; it is even more a challenging problem since users' queries are very

short (Chifu and Ionescu 2012). Indeed, more than 90% of the queries are 3 words or less long. Considering such a small number of terms, it is a challenge for the systems to "understand" the query or to disambiguate it.

To face these challenges, IR systems consider several strategies. One of them is to diversify the results by providing document related to the various senses of query terms, the system optimizes the chance of providing relevant information (Vargas et al. 2013). Another strategy is to expand the query (Gauch and Smith 1991).

The principle of query expansion (QE) is to add new query terms to the initial query in order to enhance its formulation. Candidate terms for expansion are either extracted from external resources such as WordNet or from the documents themselves; based on their links with the initial query terms. In the latter types of methods, the most popular one is the pseudo-relevance feedback (Buckley 1995).

Pseudo-relevance feedback has been shown as an effective method in average; however, it can lower results for some queries. For example, it is most probable that for poor performing queries query expansion is helpless since it will be based on the first retrieved documents that are probably non-relevant documents. It is thus important to know in advance if QE will be helpful or on the contrary if it will degrade the results. Selective query expansion aims at making this decision. However, current methods use blind methodologies and uses learning methods as black boxes. On the contrary, we think that a deep analysis of queries and query expansion terms could help understanding when QE would be useful and if there are some sort of typology of QE usefulness.

This paper pursues two objectives; first of all, we suggest a new automatic query expansion method. This new method estimates the importance of expansion candidate terms by the strength of their relation to the query terms. The second objective is to deeply analyze the results: both the initial and expanded queries and the terms they are composed of, and the cases when the expansion lowers the results and when it improves them.

The remaining of the paper is organized as follows: Section 2 reports related works. In Section 3, we present the new QE method we propose. Section 4 presents the evaluation framework. Section 5 discusses the results and presents a deep analysis of query terms, on a linguistic point of view.

2. Literature Review

2.1. QE and Pseudo Relevance Feedback

QE is either based on the analysis of a document collection (Carpineto and Romano 2012) or they imply dictionary or ontology-based methods (Bhogal, Macfarlane, and Smith 2007). This study focuses on document analysis methods. This analysis may be either (1) global (corpus analysis to detect word relationships) (Carpineto and Romano 2012) or (2) local feedback (analysis of documents retrieved by the initial query) (Rocchio 1971; Xu and Croft 1996). Local analysis or local feedback methods rely on the hypothesis that relevant documents contain terms that could be useful to enhance query formulation. Rocchio defined a method in which term weights are

re-computed so that the terms that occur in relevant documents contribute positively to the new query whereas the weight of the terms from non-relevant documents are lowered (Rocchio 1971). Rocchio's method implies to know document relevance. Buckley suggested pseudo-relevance feedback (PRF) in which the top retrieved document are automatically considered as relevant (Buckley 1995). PRF is now common practice and used in many expansion methods (Carpineto and Romano 2012). Global methods work alike but in that case candidate terms come from the entire document collection rather than just (pseudo-) relevant documents.

Divergence from Randomness (DFR) models were developed by the School of Computing Science, University of Glasgow (Ounis et al. 2006). These models are based on the assumption that informative words are relatively more frequent in relevant documents than in others. During QE the best-scored terms from the top-ranked documents are extracted. Terms are ranked using one of the DFR weighting model. DFR models include Kullback-Leibler divergence, Chi-square divergence, Bose-Einstein 1 (Bo1) and 2 (Bo2) models. In the DFR models QE is performed by ordering the candidate terms by their information content given the query Q. DFR models are presented in (G. Amati 2003). In this paper, we compare our results with Bo2.

2.2. Analysis of Queries and Results

A few studies have reported analysis of results. The deeper analysis has been conducted in the RIA Workshop that took place in 2004. One of the objectives of the workshop was to analyze the variability in systems: some systems answering well on some queries and badly on others; some other systems behaving oppositely. One of the conclusion of the workshop was that the comprehension of variability is complex because of various parameters: query formulation, the relation between the query and the documents as well as the characteristics of the system (Harman and Buckley 2009). Moreover, they conducted failure analysis for 45 of the TREC topics. After using various systems on "hard" topics, the workshop participants analyzed why the system failed. For 39 topics out of 45 the systems failed for the same reason. Moreover, even if they did not retrieve the same documents, they were missing the same aspect in the top documents. Predicting query difficulty remains a challenge (Mothe and Tanguy 2005).

The work presented in this paper combines local analysis, namely relevance feedback, and global analysis.

3. Method Description

The key idea of our new QE method is to estimate the importance of candidate terms by the strength of their relation to the query terms. In contrast to DFR models we do not compare the term frequency in PRF and the entire collection. In our approach, documents from PRF provide term candidates that are analyzed in two aspects: their frequency in PRF and their co-occurrence with query terms in the whole collection. Indeed, DFR models are based on two metrics: term frequency in PRF and

the frequency of the term in the collection. Particularly, Bo2 uses the extrapolation of term frequency in PRF on the whole collection.

In our method candidate terms are selected from the PRF and are based on the underlain hypotheses: the strength of their relation to the query terms is proportional to the fraction of the number of the documents containing both candidate terms and query terms and the product of the number of documents containing at least one of these sets.

A query is first preprocessed. It is cleared from stop-words, punctuation; duplicate terms are removed. However if a query contains only stop-words, this could mean that a user is interested, for example, in grammar. For instance, the query "a and the" may imply that a user wants to find how to use English articles. Thus, if a query contains only stop-words, we keep all of them.

The importance of term combinations w_{T_j} is estimated by the formula:

$$w_{T_j} = \sum_{t_i \in T_j} (Imp(t_i) + 1)$$

$$Imp(t_i) = \frac{1}{\log doccount(t_i)}$$

where T_j is the term combination, t_i is the i-th term from T_j, $doccount(t_i)$ is the number of documents containing the i-th term. $Imp(t_i)$ is similar to IDF. For widely-spread terms with low $Imp(t_i)$ the importance of their combination is approximately equal to their number.

The importance of candidate terms w_c is computed as follows:

$$w_c = TF(c) \times \sum_{T_j \in T} MI(T_j, c)$$

where T is the set of all possible term combinations, and $MI(T_j, c)$ is the analogue of non-negative point-wise mutual information calculated by the formula:

$$MI(T_j, c) = \frac{-\log_2 \max\left(\frac{doccount(T_j, c) \times n}{doccount(T_j) \times doccount(c)}\right)}{\log_2 \frac{doccount(T_j, c)}{n}}$$

where $doccount(c)$ is the number of documents containing the candidate term c, $doccount(T_j)$ is the number of documents containing all terms from the term combination T_j, $doccount(T_j, c)$ is the length of their intersection, and n is the total number of documents in the collection.

All weights w_c are normalized. The best-scored term candidates are selected for query expansion.

4. Evaluation

4.1. Data Collections and Evaluation Measures

The evaluation was performed on two kinds of datasets: TREC (Text Retrieval Conference) Ad Hoc Track data sets for three years (1997–1999) (Voorhees and Harman 2000) containing 150 topics in total and composed of news articles, and WT10g from TREC Web track 2000–2001 (Hawking and Craswell 2002) which is a 10GB subset of the web snapshot of 1997 from Internet Archive. There are 100 topics in this second collection from track 2000 and 2001.

In our evaluation, we considered the following evaluation measures:\begin{itemize}

- Precision at 10;
- Mean Average.

Precision (P) is the fraction of retrieved documents that are relevant. P at 10 (P@10) is the fraction of the top 10 retrieved documents that are relevant. Average precision (AP) is the average of precision computed each time a relevant document is retrieved. Mean average precision (MAP) is calculated as the mean of average precision over queries.

4.2. System Details

We compared our system (Co) with the Bo2 DFR model implemented in Terrier platform.

Both systems used InL2c1.0 model for PRF, 3 documents from which 10 best scored terms were extracted. InL2c1.0 is a DFR (divergence from randomness) model based on TF-IDF measure with L2 term frequency normalization (Gianni Amati and Van Rijsbergen 2002; He and Ounis 2005)Heidelberg",”page”:”200–214”, ”event-place”:”Berlin, Heidelberg”,”URL”:”http://dx.doi.org/10.1007/978-3-540-31865-1_15”,”DOI”:”10.1007/978-3-540-31865-1_15”,”ISBN”:”3-540-25295-9, 978-3-540-25295-5”,”author”:[{“family”:”He”,”given”:”Ben”},{“family”:”Ounis”,”given”:”Iadh”}],”issued”:{“date-parts”:[[“2005”]]}}}],”schema”:”https://github.com/citation-style-language/schema/raw/master/csl-citation.json”}.

4.3. Performance Results

Table 1 provides information about the number of relevant retrieved documents (RRD), MAP, and P@10. On adhoc data set by all metrics our systems showed the best results, which are much higher than the baseline. The Student's test confirmed that the differences between MAP values of Co and Bo2 is not significant at the level $p < 0.05$. Significant difference is marked by *, the best results are marked-up in bold. Our system showed lower results than Bo2 according to P@10. In case of web data, Bo2 obtained the same MAP score as Co. Co remains the best according to P@10.

Table 1. General results

		Co	Bo2	Baseline
TREC 6-8 data	RRD	8230	8184	7225
	MAP	0.2507	0.2491	0.2105*
	P@10	0.4400	0.4413	0.4180
Web data	RRD	3897	3935	3810
	MAP	0.2190	0.2190	0.1894*
	P@10	0.3327	0.3296	0.2816

5. Discussion and Further Analysis

5.1. Detailed Results

In the previous section, we reported the results when averaged over the set of topics. In this section, we aim at analyzing the results deeper.

Table 2. Detailed statistics

	# topics	# Baseline best	# Bo2 best	# Co best
Adhoc	150	33 (22%)	52 (34.6%)	63 (42%)
Adhoc 25 hardest	25	8 (32%)	7 (28%)	10 (40%)
Adhoc 25 easiest	25	7 (28%)	10 (40%)	7 (28%)
Web	98	34 (34.7%)	30 (30.6%)	31 (31.6%)
Web 25 hardest	25	11 (44%)	5 (20%)	7 (28%)
Web 25 easiest	25	10 (40%)	10 (40%)	4 (16%)

Table 2 reports the number of topics in each collection for which the method (column) got the best results according to AP. We also report these numbers when the 25 hardest and 25 easiest topics are considered. The hardest and easiest topics are defined as the ones that got the highest and lowest AP using the initial query. For example, in the Adhoc collection, from 150 topics, 33 are best treated without QE, 52 best when using Bo2 and 63 when using our method. The percentage of best treated topic per method is slightly different when considering the easiest topics: 7 are best without QE, 10 are best treated using Bo2 and 7 using our method.

Two examples are provided below. Each topic is composed of the title part which was used as a submitted query to the system, as well as descriptive and narrative parts that helps in understanding the user's need.

Example 1
<num>530
<title>do pheromone scents work?
<desc>What is the scientific evidence that suggests pheromones stimulate the opposite sex?
<narr>A relevant document will discuss how pheromones act as an attractor or repellent among humans, other animals, or plants.

Example 2.
<num>494
<title>nirvana
<desc>Find information on members of the rock group Nirvana.
<narr>Descriptions of members' behavior at various concerts and their performing style is relevant. Information on who wrote certain songs or a band member's role in producing a song is relevant. Biographical information on members is also relevant.

Table 3 provides the results in terms of AP and the various query formulations the system really used as well as the term weights. For example, with regard to topic 530, "do" was removed as a stop word and the other words have been stemmed.

Table 3. Query reformulation examples

# Topic	AP / Initial query	AP / Bo2 reformulation	AP / Co reformulation
530	0.3838 / pheromon^ 1.0 scent^ 1.0 work^ 1.0	0.1551/ pheromon^1.60 scent^ 1.30 work^1.00 fragranc^ 1.00 perfum^0.96 men^ 0.45 attract^ 0.41 design^ 0.41 sex^ 0.38 natur^ 0.34 sexual^ 0.34	0.1765 / pheromon^ 2.0 scent^ 1.68 work^ 1.42perfum^ 1.0fragranc^ 0.97 aphrodisiac^ 0.36 sex^ 0.22 men^ 0.22 attract^ 0.21 sexual^ 0.17 world^ 0.16 cologn^ 0.15 natur^ 0.15 nerd^ 0.13
494	0.1706 / nirvana^ 1.0	0.3508 / nirvana^ 2.00 kurt^0.29 cobain^ 0.25 bootleg^0.17 world^ 0.17 list^0.15 song^ 0.13 new^0.11 contain^ 0.11 dedic^0.11	0.5703 / nirvana^2.0 cobain^ 1.0 kurt^0.82 bootleg^ 0.42 song^0.30 nerd^ 0.26 music^ 0.18 unplug^0.18 band^ 0.17 world^0.16 sound^ 0.15 stuff^0.14

In topic 530 both reformulation went to the "sex" concept which leaded to some noise in the answers.

In topic 494, it is clear that adding “kurt” and “cobain”, the lead singer, guitarist, and primary songwriter of the band Nirvana help in retrieving relevant documents. Either the weight of the terms (which are stronger in our reformulation than in Bo2) or some additional terms such as “band” made our reformulation better than Bo2.

5.2. Types of Initial Queries

Types of initial queries play an essential role in the prediction of successful information retrieval. As usual initial queries include, besides articles and other grammar words, nouns and entities, sometimes attributes and verbs. Grammatical structure of a title does not influence on the information retrieval process, because every title while processing the query is ruined into words and even word chunks. So types of queries are limited by a number of words and topic. Types of the query terms are restricted to words grammatical classes, such as parts of speech, and words semantic classes, such as terminology, entities, peculiarities, etc.

Potentially a document matches the initial query thanks to one term, or one term with its attribute, or two (or more) different terms. The last possibility is the best one, since a number of documents with two (or more) unconnected terms from the initial query is less, than a number of documents with the term and its attribute (noun phrase). In other words, co-occurrence of two (or more) semantically unconnected query terms in a document guarantees more accurate matching the initial query, while occurrence of one term just presupposes matching in a topic. Thus, one term query is less informative, than two and more terms queries. Hence for a one-word initial query QE is a productive way to increase the relevance of results, however, it depends on semantics of the one-word query. Our results for QE for one-word queries are slightly better regardless of the QE methods.

As a consequence of the diffusive character of the category, there are a lot of different factors which influence on the document frequency of the words associated with the topic. Thus, the more texts we use for the QE in the global analysis, the more unpredictable candidates we get for the QE. That is why the query generated on the basis of the title *What is a Bengals Cat?* (Animals) provides slightly better results with Bo2 QE system, while, with our system, we get a lot of useless extensions.

The structure of scientific categories is more compact and hierarchical. We assume that the initial query in the field of scientific categories evokes texts with less associative and more logical connections. The QE allows directing the IR process in a narrow relevant field. The title *Unsolicited Faxes* (Ad Hoc) refers to a multi-topic document, which simultaneously belongs to at least two topics in our set (“crimes” and “technology”). The results of QE performed by both systems are very good, but for our system it is significantly better: 0.6638 against 0.6015.

Therefore, the topic of the initial query is a strong factor, which influences on the necessity of the QE. Within homogenous text collection, every QE system works good, producing better results, than an initial query. Within naive topics categories the simple QE system is appropriate, while our QE generates complicated associative queries. So for the IR on the topic from naive category within heterogeneous text collection our QE system is overcomplicated, and that is why it works worse.

6. Conclusion

In this paper, we first suggest a new method for QE we call Co. The key idea of the proposed method is to estimate the importance of candidate terms by the strength of their relation to the query terms.

In our experiments, we show that the Co method has similar results as the Bo2 method from the literature. However, our finer analysis shows that the type of initial query can have an influence on the success of QE.

In our future work, we will work on the relationship between the types of queries and the field associated to the query in order to detect correlation with these features and the best method to treat the query. We think that using more linguistic features can help in selective approaches in IR.

References

1. *Amati, G.* (2003), Probability Models for Information Retrieval Based on Divergence from Randomness: PhD Thesis. University of Glasgow.
2. *Amati, Gianni, and Cornelis Joost Van Rijsbergen* (2002), Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. ACM Trans. Inf. Syst. Vol. 20 no. 4 pp. 357–389 (October).
3. *Bhogal, J., A. Macfarlane, and P. Smith* (2007), A Review of Ontology Based Query Expansion. Inf. Process. Manage. Vol. 43 no. 4 pp. 866–886 (July).
4. *Buckley, Chris* (1995), Automatic Query Expansion Using SMART : TREC 3. In Proceedings of The Third Text REtrieval Conference (TREC-3). NIST Special Publication 500–226. pp. 69–80. Gaithersburg, MD: National Institute of Standards and Technology (NIST).
5. *Carpineto, Claudio, and Giovanni Romano* (2012), A Survey of Automatic Query Expansion in Information Retrieval. ACM Computing Surveys. Vol. 44 no. 1 pp. 1–50 (January).
6. *Chifu, Adrian-Gabriel, and Radu-Tudor Ionescu* (2012), Word Sense Disambiguation to Improve Precision for Ambiguous Queries. Cent. Eur. J. Comp. Sci. no. in printno. in print.
7. *Gauch, Susan, and John B. Smith* (1991), Search Improvement via Automatic Query Reformulation. ACM Trans. Inf. Syst. Vol. 9no. 3pp. 249–280.
8. *Harman, Donna, and Chris Buckley* (2009), Overview of the Reliable Information Access Workshop. Information Retrieval. Vol. 12 no. 6 pp. 615–641 (July).
9. *Hawking, David, and Nick Craswell* (2002), Overview of the TREC-2001 Web Track. NIST Special Publication. pp. 61–67.
10. *He, Ben, and Iadh Ounis* (2005), Term Frequency Normalisation Tuning for BM25 and DFR Models. In Proceedings of the 27th European Conference on Advances in Information Retrieval Research. pp. 200–214. ECIR'05. Berlin, Heidelberg: Springer-Verlag. http://dx.doi.org/10.1007/978-3-540-31865-1_15.

11. *Mothe, J., and L. Tanguy* (2005), Linguistic Features to Predict Query Difficulty—A Case Study on Previous TREC Campaign. SIGIR Workshop on Predicting Query Difficulty—Methods and Applications.
12. *Ounis, I., G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma* (2006), Terrier: A High Performance and Scalable Information Retrieval Platform. In Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006). Seattle, Washington, USA.
13. *Rocchio, J.* (1971), Relevance Feedback in Information Retrieval. In The SMART Retrieval System, 313–323. http://scholar.google.com/scholar?hl=en&lr=&client=firefox-a&q=relevance+feedback+in+information+retrieval&btnG=Search.
14. *Vargas, S., R. L. T. Santos, C. Macdonald, and I. Ounis* (2013), Selecting Effective Expansion Terms for Diversity. In 10th International Conference in the RIAO Series (OAIR 2013). Lisbon, Portugal. http://ir.ii.uam.es/predict/pubs/oair2013-vargas-gla.pdf.
15. *Voorhees, Ellen M., and Donna Harman* (2000), Overview of the Ninth Text REtrieval Conference (TREC-9). In Proceedings of the Ninth Text REtrieval Conference (TREC-9). pp. 1–14.
16. *Xu, Jinxi, and W. Bruce Croft* (1996), Query Expansion Using Local and Global Document Analysis. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 4–11. SIGIR'96. New York, NY, USA: ACM. http://doi.acm.org/10.1145/243199.243202.

РАБОЧАЯ ПАМЯТЬ И РУССКИЙ ЯЗЫК: ОТ РЕЧЕПОНИМАНИЯ К РЕЧЕПОРОЖДЕНИЮ

Федорова О. В. (olga.fedorova@msu.ru),
Потанина Ю. Д. (binechka-paveletskaja@mail.ru)

Московский государственный университет
им. М. В. Ломоносова, Москва, Россия

Ключевые слова: рабочая память, психолингвистика, понимание речи, порождение речи

WORKING MEMORY AND RUSSIAN LANGUAGE: FROM COMPREHENSION TO PRODUCTION

Fedorova O. V. (olga.fedorova@msu.ru),
Potanina Ju. D. (binechka-paveletskaja@mail.ru)

Lomonosov Moscow State University, Moscow, Russia

Working memory and long-term memory differ in many ways. One difference is in the storage capacity of each. Traditionally, the capacity of the working memory has been measured by a memory span task in which the individual hears series of items and must repeat them. Most of the research has focus on individual differences in working memory capacities. Daneman and Carpenter (1980) developed the Reading span test, which they interpret as providing a measure of an individual's working memory capacity. The subject is given a series of sentences to read, and then must recall the last word from each of the preceding sentences. Span is calculated as the maximum number of sentences on which the subject can perform this task perfectly. In 1986 Daneman and Green developed the Speaking span test. Most of the research has done on English-speaking individuals. The main goal of this paper is to provide and describe the Verbal span tests on Russian material. The present study shows how the use of the notion of verbal working memory contributes to our understanding the individual differences in language comprehension and language production mechanisms. Using Russian adaptations of the working memory reading span and speaking span tests we demonstrated that the working memory capacity is really correlated with some referential processes, as well as it is a predictor of verbal fluency.

Key words: working memory, psycholinguistics, language comprehension, language production

Введение[1]

Когнитивные психологи занимаются изучением рабочей памяти (далее РП) уже более ста лет (см. раздел 1), последние 35 лет им в этом активно помогают психолингвисты, первые исследования которых датируются 1980-ми гг. XX в. До недавнего времени подавляющее большинство психолингвистических работ проводилось на материале английского языка. Десять лет назад, только начиная заниматься изучением взаимодействия рабочей памяти и языка, мы представили на конференции «Диалог» доклад, посвященный разработке первого **русскогоязычного** теста по определению объема РП при **понимании речи** (Федорова 2003). За прошедшие десять лет мы провели несколько серий подобных экспериментов (раздел 2). Раздел 3 будет посвящен описанию взаимодействия РП и механизмов **порождения речи**, исследования в области изучения которых на русском материале начались совсем недавно. В каждом из двух последних разделов сначала мы опишем методику определения объема РП, а потом приведем примеры русскоязычных исследований, показывающих корреляцию между языковым поведением испытуемых и объемом их РП. В заключении будут кратко описаны перспективы дальнейших исследований.

1. Исследования памяти в когнитивной психологии

Начало современного этапа изучения памяти традиционно датируется концом XIX в. и связывается с именем Г. Эббингауза, разработавшего первые экспериментальные методики, с помощью которых ему удалось определить важные закономерности функционирования памяти (Ebbinghaus 1885); 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), Н. Во и Д. Нормана (Waugh, Norman 1965) и Р. Аткинсона и Р. Шиффрина (Atkinson, Shiffrin 1968) был сформулирован многокомпонентный подход к памяти. Согласно модели Аткинсона и Шиффрина (1968), сначала информация попадает в сенсорные регистры, затем переводится в кратковременное хранилище, после чего передается в долговременную память. Термин ‘рабочая память’ (англ. working memory) был впервые использован в книге Miller et al. 1960. В то время как в модели Аткинсона и Шиффрина термином РП обозначался атомарный блок трехкомпонентной модели памяти, А. Бэддели и Дж. Хитч (1974) использовали его уже для обозначения некоторой сущности, которая сама состоит из трех отдельных компонентов. Употребление термина РП вместо более традиционного термина ‘кратковременная память’ подчеркивает функциональную важность системы, то есть авторы,

[1] Исследование проведено при частичной поддержке грантов РГНФ 14-04-00390 и РФФИ 14-06-00211.

использующие этот термин, в первую очередь ищут ответ на вопрос, для каких целей служит РП.

Трехкомпонентная модель РП Бэддели и Хитча (1974) состоит из центрального исполнителя, фонологической петли и визуально-пространственной матрицы. Центральный исполнитель является ядром системы, отвечающим за координацию работы подсистем, а два 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, без поддержки которого информация в фонологическом хранилище угасает примерно через 1,5 с. (Baddeley et al. 1975). Зрительная информация попадает из сенсорного хранилища в визуально-пространственную матрицу, которая состоит из зрительной и пространственной подсистем.

В настоящее время существует много подтверждений психологической обоснованности выделения фонологической петли и визуально-пространственной матрицы в отдельные блоки. Например, в эксперименте, описанном в Baddeley et al. 1975, испытуемых просили прочитать и запомнить ряд названий различных государств, причем в одной группе названия были короткими (*Чад, Кения, Чили*), а в другой — длинными (*Чехословакия, Швейцария, Австралия*). Оказалось, что испытуемые из первой группы лучше справляются с заданием, чем испытуемые из второй. Авторы объясняют этот эффект тем, что в процессе запоминания человек обычно проговаривает слова про себя. Чем длиннее слово в последовательности, тем меньше раз испытуемые успевают проговорить эту последовательность за отведенное время, и тем дольше длится время угасания следа от предыдущих проговоренных слов.

В 2000 г. в трехкомпонентную модель РП был добавлен еще один компонент — эпизодический буфер, который используется для синтеза и интеграции информации из фонологической петли и визуально-пространственной матрицы, а также для связи с долговременной памятью (Baddeley 2000). Приведем один из аргументов, побудивший Бэддели добавить в свою модель этот компонент. Давно известно, что человек может удерживать в памяти последовательность из пяти-семи слов, не связанных между собой. Однако если эти слова объединить в предложения, то количество слов, которые человек может запомнить, возрастает в несколько раз. Этот пример является хорошей иллюстрацией явления, которое Миллер (1956) назвал структурированием. Однако где хранятся эти структурированные группы слов? Бэддели предполагает, что такая собранная из разных источников информация хранится как раз в эпизодическом буфере. Итак, в отличие от прежних моделей, в модели Бэддели утверждается, что поступающая в РП информация не только пассивно хранится, но и активно обрабатывается. Подобная идея проходит лейтмотивом по всем исследованиям РП последних десятилетий.

2. Исследования рабочей памяти в психолингвистике: понимание речи

Вскоре после выхода работы Бэддели и Хича (1974) термин РП был впервые использован в психолингвистической статье Daneman, Carpenter 1980. В конце 70-ых гг. XX в. среди исследователей процессов понимания речи сложилась парадоксальная ситуация, когда интуитивно они были уверены, что индивидуальные различия в объеме РП должны оказывать влияние на механизмы понимания речи, однако проводимые ими эксперименты этого не подтверждали. Данеман и Карпентер (1980) предположили, что все дело в отсутствии адекватных методик определения объема кратковременной памяти и предложили новый тест, которому суждено было совершить переворот в современной психолингвистике.

Данеман и Карпентер (1980) исходили из того, что в процессе интерпретации некоторого текста в РП человека происходят процессы, связанные как с пассивным хранением поступающей информации, так и с ее обработкой. Существовавшие же в то время методики по определению объема кратковременной памяти (состоявшие в запоминании отдельных цифр и/или слов) тестировали только первую из этих двух составляющих, нивелируя тем самым индивидуальные различия испытуемых; между тем эти различия и возникают, по мнению авторов, вследствие лучшей или худшей способности испытуемых эффективно распределять имеющиеся ресурсы РП, отводя какую-то часть для хранения поступающей информации, а другую — для ее обработки. Другими словами, чем меньше ресурсов затрачивается на обработку поступающей информации, тем больше их остается для ее хранения. Тест, получивший название **Reading span**, тестировал обе этих составляющих — в ходе эксперимента испытуемый должен был читать отдельные предложения и одновременно удерживать в РП последние слова ранее прочитанных предложений. Таким образом, по словам Данеман, «теория кратковременной памяти была заменена теорией РП, а методика измерения кратковременной памяти — методикой измерения РП» (Daneman 1994: 443).

Ниже мы опишем современную версию данного теста, которая, впрочем, совсем немного отличается от оригинальной. Тест состоит из 70 предложений, взятых из литературных источников. Каждое предложение напечатано в центре небольшой карточки, которые распределены на группы по 2, 3, 4 и 5 предложений, по пять попыток на каждом уровне. В ходе эксперимента экспериментатор выкладывает перед испытуемым по одной карточке и просит его прочитать написанное на ней предложение вслух. Как только испытуемый заканчивает чтение предложения, экспериментатор накрывает эту карточку следующей. После того как экспериментатор кладет перед испытуемым пустую карточку, тот должен повторить последние слова каждого предложения этой группы в том порядке, в котором они были им прочитаны, с точностью до словоформы. Эксперимент продолжается до тех пор, пока испытуемый может воспроизвести не меньше трех из пяти групп на данном уровне. В противном случае эксперимент заканчивается, и объем РП испытуемого считается равным последнему уровню, на котором он смог воспроизвести последние слова трех из пяти предложенных ему предложений.

Русскоязычная версия данного теста была разработана в 2001 году, см. Федорова 2003; Fedorova, Pechenkova 2007; Федорова, Потанина 2013). За прошедшее в тех пор время было проведено более полутора тысяч подобных экспериментов, по результатам которых примерно 65% русскоязычных испытуемых имеют небольшой объем РП, что значимо отличается от англоязычных исследований, в которых эта цифра закрепилась на уровне 50%. Основной причиной такого положения дел, на наш взгляд, является богатое русское словоизменение, которое сильно увеличивает количество ошибок при повторении словоформ (Fedorova, Pechenkova 2007).

Для иллюстрации зависимости речевого поведения человека от объема его РП при понимании речи приведем пример из области дискурса. В Daneman, Carpenter 1980 испытуемые читали небольшие дискурсивные фрагменты и должны были правильно установить антецедент местоимения *he*, которое было использовано в последнем предложении. При этом первое упоминание целевого референта встречалось в дискурсивных отрывках на разном линейном расстоянии от местоимения, их разделяло от 2 до 7 предложений. Авторы убедительно показали, что испытуемые с большим объемом РП выполняют это задание значительно лучше, чем испытуемые с небольшим объемом.

В Fedorova et al. 2010 описан аналогичный эксперимент, в котором варьируется **риторическое расстояние** до антецедента местоимения, традиционно вычисляемое по иерархической структуре, разработанной в рамках Теории риторической структуры (Mann, Thompson 1988). Испытуемые читали текст (см. пример 1), а затем отвечали на три вопроса, из которых первые два были фактографические, а третий — референциальный:

(1) *Был конец рабочего дня. Пятая бригада скорой помощи ехала на базу после ложного вызова. На носилках в кабине, набегавшись за смену, прикорнул медбрат. Усталый* ***доктор****, слушавший музыку в плеере, игнорировал заискивающие взгляды молодого ассистента, горящего рабочим энтузиазмом после первого дня в бригаде. В наушниках звучал «Белый альбом» битлов. Безупречная мелодия качала и убаюкивала.*
а) РитР=1 ***Он*** *испытывал легкие угрызения совести за свою невнимательность к коллеге, но усталость превозмогала всё.*
б) РитР=2 ***Он*** *любил слушать эту пластинку после тяжелого трудового дня.*
в) РитР=3 ***Он*** *почувствовал, что медленно проваливается в сон.*

Вопросы*:*
1. *Какой номер был у бригады скорой?*
2. *Какая запись звучала в плеере?*
3. а) *Кому было стыдно за невнимательность к коллеге?*
   б) *Кто любил слушать пластинку после тяжелого трудового дня?*
   в) *Кто почувствовал, что засыпает?*

Проведя эксперимент с 120 испытуемыми, мы обнаружили, что чем выше объем РП испытуемых, тем меньше ошибок они совершают в ответах на референциальные вопросы ($cor = -0{,}56$; $p\text{-value} < 0{,}01$). Таким образом, мы

показали, что успешность интерпретации референциального выражения напрямую зависит от объема РП испытуемых. Эффект зависимости механизмов речепонимания от объема РП был обнаружен нами и при анализе сложноподчиненных предложений с относительными придаточными (Fedorova, Yanovich 2006), а также с придаточными времени (Федорова 2005; Fedorova 2010). Подытоживая вышеизложенное, можно заключить, что в области изучения понимания речи зависимость языкового поведения человека от объема его РП является неоспоримым фактом, многократно подтвержденным в исследованиях. Иначе обстоит дело в области порождения речи, где данный факт еще только предстоит строго доказать.

3. Исследования рабочей памяти в психолингвистике: порождение речи

В конце 80-ых гг. XX века один из авторов статьи Daneman, Carpenter 1980 — М. Данеман — разработала новый тест на определение объема РП, связанный с порождением речи. Данный тест получил название **Speaking span** (Daneman, Green 1986; Daneman 1991). Для эксперимента было отобрано 100 слов, которые были распределены в группы по 2, 3, 4, 5 и 6 слов. Каждое слово появлялось на экране на 1 с.; испытуемый получал инструкцию читать слова и, увидев пустой экран, придумывать с каждым прочитанным словом по одному предложению, причем целевое слово в этом предложении должно было стоять в той же форме. Например, прочитав слова *shelter, muscles* и *dangers,* англоязычный испытуемый произносит предложения: *Trees provide poor **shelter** during a thunderstorm*; *Mr. Universe has very big **muscles***; *There are **dangers** associated with every occupation*. Испытуемым разрешалось придумывать предложения в любом порядке, но было запрещено использовать последнее прочитанное слово первым.

В результате объем РП условно приравнивался к количеству слов, с которыми испытуемый мог придумать предложения. В работе (Daneman 1991) автор продемонстрировала, что объем РП, измеренный с помощью теста Speaking span, коррелирует с **беглостью речи** при порождении. Настоящий раздел посвящен проверке гипотезы о взаимозависимости объема РП и беглости речи на материале русского языка. Эксперимент, проведенный с 32 испытуемыми, состоял из пяти тестов: 1) Speaking span; 2) порождение речи; 3) чтение вслух; 4) оговорки; 5) чтение скороговорок. Испытуемые в основном являлись студентами филологического факультета МГУ им. Ломоносова в возрасте от 18 до 25 лет.

При создании русскоязычной версии теста Speaking span были использованы следующие ограничения: все слова были семибуквенными, высокочастотными, сбалансированными по частеречной принадлежности, а также равномерно распределенными по грамматическим признакам в соответствии с частотностью употребления грамматической формы; кроме того, между словами в группе нельзя было установить ассоциативные связи. Объем РП подсчитывался в процентах и по шкале от 2 до 6.

Для теста на порождение речи была выбрана фотография семьи за обедом. Испытуемых просили описывать фотографию в течение одной минуты, мерой беглости речи считалось общее количество произнесенных слов.

В ходе теста на чтение вслух испытуемых просили прочитать отрывок длиной в 328 слов («Подросток» Ф. М. Достоевского), как можно быстрее и четче произнося слова. При обработке результатов для каждого испытуемого было посчитано число ошибок и время, за которое он прочитывал весь текст. Ошибками считались повторы, фальстарты, оговорки, добавления, пропуски и замены.

Наиболее трудоемким оказалась разработка теста на оговорки. Процедура эксперимента повторяла эксперимент из Daneman 1991 и состояла в следующем: на экране компьютера предъявлялись 309 пар слов, по 1 с. на каждую пару. Испытуемые читали все пары слов про себя за исключением определенных пар (маркированных звуковым сигналом), которые они читали вслух. 30 экспериментальных пар были подобраны таким образом, чтобы вызвать оговорку; кроме того, в эксперименте было 39 филлерных пар, необходимых для того, чтобы скрыть от испытуемого реальную цель эксперимента. Оговорка провоцировалась тремя парами фонологически похожих слов, например: *суетные мысли, сушит мышцы, сунул мыло,* ***мушки сыты***. Первые три пары слов похожи на ожидаемую оговорку ***сушки мыты*** — они имеют аналогичную ритмическую структуру и одинаковые начальные звуки. Кроме того, ожидаемая оговорка представляла собой реально возможное словосочетание.

Для эксперимента со скороговорками были отобраны 15 русских скороговорок. Их длина варьировалась от 6 до 24 слов. Для того чтобы нивелировать индивидуальные различия между испытуемыми (например, сложности с произнесением определенных звуков) подбирались скороговорки с разными звуками и их сочетаниями ([б], [в], [г], [ж], [з], [к], [л], [м], [н], [п], [р], [с], [т], [ф], [ц], [ч], [ш], [щ], [кл], [кр], [гр], [тр]). В стимульный материал не отбирались слишком длинные скороговорки, содержащие многосложные слова, так как на такие предложения испытуемому могло бы не хватить запаса воздуха в легких, что привело бы к длительным паузам, связанным с физиологическими особенностями речепорождения. Порядок предъявления скороговорок был произвольным. Для проведения эксперимента использовалась программа PowerPoint, в которой испытуемым последовательно (по одной на слайд) предъявлялись скороговорки. После прочтения скороговорки вслух испытуемый нажимал на клавишу на клавиатуре, и на экран выводилась новая скороговорка. Эксперимент записывался на диктофон, фиксировалось количество оговорок и время чтения. После прочтения всей группы скороговорок участникам предлагается отдохнуть. На следующем этапе эксперимента испытуемым давалось 5 минут для того, чтобы подготовиться ко второму подходу к чтению. За эти 5 минут участники эксперимента могли перечитывать все скороговорки в любом порядке, тренироваться произносить их вслух, заучивать, экспериментатор не ограничивал их в методах самоподготовки. По истечении 5 минут, мы снова просили испытуемых как можно быстрее прочитать скороговорки и опять записывали их на диктофон, фиксируя количество ошибок и время чтения. Мерами беглости речи в этом эксперименте считались 1) время чтения и 2) количество оговорок.

В результате мы получили значимые корреляции между объемом РП и (1) количеством слов в тесте на порождение речи (cor = 0,522, p-value < 0,05); (2) временем чтения вслух (cor = −0,704, p-value < 0,01); (3) тестом на оговорки (cor = −0,706, p-value < 0,01); (4) временем чтения скороговорок (cor = −0,500 и −0,471 для первой и второй попыток, соответственно, p-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. Тем не менее, результаты в целом показывают, что объем РП является значимым фактором, коррелирующим с беглостью речи русскоязычных испытуемых.

Заключение

В настоящей работе мы показали, что объем РП коррелирует с индивидуальными различиями людей в способности понимать обращенную к ним речь, а также с их способностью к порождению речи. Более того, в работах классиков данного направления была выдвинута идея о делении всех людей на «хороших» читателей, умеющих эффективно распределять ресурсы РП между хранением и обработкой поступающей информации, и «плохих» читателей, которые делают это хуже (Daneman, Carpenter 1980). Хотя этот последний тезис не находит подтверждения во многих современных работах (см., например, Otten, van Berkum 2009), большая роль РП при речепонимании несомненна. В то же время, несмотря на то, что еще в 1980 году А. Эллис заметил, что ошибки, которые совершаются в процессе прохождения теста по определению объема РП, аналогичны ошибкам, совершаемым людьми при порождении в обычной повседневной жизни (Ellis 1980), корреляции между объемом РП и порождением речи исследованы еще недостаточно хорошо. Данная тенденция повторяет общий тренд психолингвистической науки уделять значительно больше внимания изучению процессов понимания речи, чем процессам речепорождения. Однако в последние годы ситуация начинает меняться, что проявляется, в частности, в привлечении к исследованиям речепорождения более широкого, когнитивно ориентированного контекста, лежащего в пограничной области между «чистой» психолингвистикой и когнитивной психологией. Кроме того, отдельной составляющей дальнейшего изучения данной проблематики является вопрос о корреляции между двумя описанными выше тестами: если испытуемый хорошо справился с тестом по определению объема РП при понимании речи, можем ли мы что-то предсказать относительно его результатов в тесте на порождение речи, и наоборот?

Литература

1. *Федорова О. В.* (2003), Тест по определению объема оперативной памяти: история и современное состояние, Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2003», М.

2. *Федорова О. В.* (2005), Перед или после: что проще?, Вопросы языкознания, 6.
3. *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.
4. *Baddeley A. D., Hitch G. H.* (1974), Working memory, G. A. Bower (ed.) Recent advances in learning and motivation, New York.
5. *Baddeley A. D., Thomson N., Buchanan M.* (1975), Word length and the structure of short-term memory, Journal of Verbal Learning and Verbal Behavior, 14.
6. *Baddeley A. D.* (2000), The episodic buffer: A new component of working memory?, Trends in Cognitive Sciences, 4.
7. *Daneman M., Carpenter P. A.* (1980), Individual differences in working memory and reading, Journal of Verbal Learning and Verbal Behavior, 19.
8. *Daneman M., Green I.* (1986), Individual differences in comprehending and producing words in context, Journal of Memory and Language, 25.
9. *Daneman M.* (1991), Working memory as a predictor of verbal fluency, Journal of Psycholinguistic Research, 20.
10. *Daneman M.* (1994), Working memory and language, Language and Speech, 37.
11. *Ebbinghaus H.* (1885), Über das Gedächtnis.
12. *Ellis A. W.* (1980), Errors in speech and short-term memory: The effects of phonemic similarity and syllable position, Journal of Verbal Learning and Verbal Behavior, 19.
13. *Fedorova O. V., Yanovich I. S.* (2006), Early preferences in relative clause attachment in russian: The effect of working memory differences, J. E. Lavine et al. (eds.) Formal Approaches to Slavic Linguistics, Ann Arbor.
14. *Fedorova O. V., Pechenkova E. V.* (2007), When «Colorless green ideas…» meet working memory span, V. D. Solovyev, E. V. Pechenkova, V. N. Polyakov (eds.) Proceedings of the 9th International conference «Cognitive modeling in linguistics», Kazan'.
15. *Fedorova O. V., Delikishkina E. A., Uspenskaya A. M.* (2010), Experimental approach to reference in discourse: Working memory capacity and language comprehension in Russian, Proceedings of the Pacific Asia Conference on Language, Information and Computation. Sendai.
16. *Fedorova O. V.* (2010), Why the English easiest type became the hardest in Russian, or Russian adults' comprehension of before and after sentences, Proceedings of the Pacific Asia Conference on Language, Information and Computation. Sendai.
17. *James W.* (1890), The principles of psychology, New York.
18. *Mann W., Thompson S. A.* (1988), Rhetorical structure theory: Toward a functional theory of text organization, Text, 8.
19. *Miller G. A.* (1956), The magical number seven, plus or minus two: Some limits on our capacity for processing information, Psychological Review, 63.
20. *Miller G. A., Galanter E., Pribram K. H.* (1960), Plans and the structure of behavior, New York.
21. *Otten M., van Berkum J.* (2009), Does working memory capacity affect the ability to predict upcoming words in discourse?, Brain research, 1291.
22. *Waugh N. C., Norman D. A.* (1965), Primary memory, Psychological Review, 72.

References

1. *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.
2. *Fedorova O. V.* (2005), Pered or posle: What is easy? [Pered ili posle: Chto proshche?], Voprosy Jazykoznanija [Voprosy Jazykoznanija], 6.
3. *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.
4. *Baddeley A. D., Hitch G. H.* (1974), Working memory, G. A. Bower (ed.) Recent advances in learning and motivation, New York.
5. *Baddeley A. D., Thomson N., Buchanan M.* (1975), Word length and the structure of short-term memory, Journal of Verbal Learning and Verbal Behavior, 14.
6. *Baddeley A. D.* (2000), The episodic buffer: A new component of working memory?, Trends in Cognitive Sciences, 4.
7. *Daneman M., Carpenter P. A.* (1980), Individual differences in working memory and reading, Journal of Verbal Learning and Verbal Behavior, 19.
8. *Daneman M., Green I.* (1986), Individual differences in comprehending and producing words in context, Journal of Memory and Language, 25.
9. *Daneman M.* (1991), Working memory as a predictor of verbal fluency, Journal of Psycholinguistic Research, 20.
10. *Daneman M.* (1994), Working memory and language, Language and Speech, 37.
11. *Ebbinghaus H.* (1885), Über das Gedächtnis.
12. *Ellis A. W.* (1980), Errors in speech and short-term memory: The effects of phonemic similarity and syllable position, Journal of Verbal Learning and Verbal Behavior, 19.
13. *Fedorova O. V., Yanovich I. S.* (2006), Early preferences in relative clause attachment in russian: The effect of working memory differences, J. E. Lavine et al. (eds.) Formal Approaches to Slavic Linguistics, Ann Arbor.
14. *Fedorova O. V., Pechenkova E. V.* (2007), When "Colorless green ideas..." meet working memory span, V. D. Solovyev, E. V. Pechenkova, V. N. Polyakov (eds.) Proceedings of the 9th International conference "Cognitive modeling in linguistics", Kazan'.
15. *Fedorova O. V., Delikishkina E. A., Uspenskaya A. M.* (2010), Experimental approach to reference in discourse: Working memory capacity and language comprehension in Russian, Proceedings of the Pacific Asia Conference on Language, Information and Computation. Sendai.
16. *Fedorova O. V.* (2010), Why the English easiest type became the hardest in Russian, or Russian adults' comprehension of before and after sentences, Proceedings of the Pacific Asia Conference on Language, Information and Computation. Sendai.
17. *James W.* (1890), The principles of psychology, New York.

18. *Mann W., Thompson S. A.* (1988), Rhetorical structure theory: Toward a functional theory of text organization, Text, 8.
19. *Miller G. A.* (1956), The magical number seven, plus or minus two: Some limits on our capacity for processing information, Psychological Review, 63.
20. *Miller G. A., Galanter E., Pribram K. H.* (1960), Plans and the structure of behavior, New York.
21. *Otten M., van Berkum J.* (2009), Does working memory capacity affect the ability to predict upcoming words in discourse?, Brain research, 1291.
22. *Waugh N. C., Norman D. A.* (1965), Primary memory, Psychological Review, 72.

КОЛЬЦО И *ЩЕПОТЬ*: СЕМАНТИКА СОЕДИНЕННЫХ ПАЛЬЦЕВ В РУССКОЙ ЖЕСТИКУЛЯЦИИ[1]

Гришина Е. А. (rudi2007@yandex.ru)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

В статье проанализированы основные русские жесты, включающие в себя в качестве компонента соединение пальцев (конфигурации *точь-в-точь, перо, щепоть*), и определены те лингвистические значения, которые этими жестами передаются (‘точность’, ‘маленький объект’, ‘объект’, ‘центр’, ‘соединение’).

Ключевые слова: соотношение слова и жеста; русская жестикуляция; конфигурация рук; Мультимедийный русский корпус (МУРКО)

RING AND *GRAPPOLO*: FINGERTIP CONNECTIONS IN RUSSIAN GESTICULATION AND THEIR MEANINGS

Grishina E. A. (rudi2007@yandex.ru)

Institute of Russian Language RAS, Moscow, Russia

The study analyzes the main types of Russian gestures, which are based on the connection of one’s fingertips (configuration *exactly, feather, bunch*). We distinguish five semantic groups, which correspond to these configurations (‘exactness’, ‘small object’, ‘object’, ‘center’, ‘connection’). We also compare the linguistic functions of the fingertip connections and the hand physical contact.

Key words: gesture-word coordination; hand configurations; Russian gesticulation; Multimodal Russian Corpus (MURCO)

[1] Исследование проведено при поддержке программы Президиума РАН «Корпусная лингвистика» (2012–2014 гг.) и гранта РФФИ № 14-06-00245.

1. Введение

В данной статье на материале некинематографической зоны Мультимедийного русского корпуса (МУРКО)[2] будут проанализированы основные значения, которые имеют в русской жестикуляции конфигурации рук, в качестве основного компонента включающие в себя соединение пальцев. Две основные конфигурации такого типа уже неоднократно были предметом исследований, прежде всего, на материале итальянского языка. Этими конфигурациями являются *кольцо* (*ring*, см. [de Jorio 1832/2000], [Munari 1963], [Efron 1972], [Morris et al., 1979], [Diadori 1990], [Kendon 2004], [Calbris 2011]) и *щепоть* (*purse-hand* рука-кошелек [Morris et al. 1979], *finger bunch* связка пальцев [Kendon 1995], итал. *grappolo* связка, кластер [Kendon 2004], *pyramid* пирамида [Calbris 2011]). Ж. Кальбрис на французском материале различает два типа кольца — *finger-nail pincers* (*букв*. пинцет, щипцы, образованные ногтями) и *finger pinch* (*букв*. щепотка, щипок пальцами).

Описать значение этих конфигураций рук можно цитатой из [Kendon 2004:236, 238, 240, *перевод и шрифтовые выделения мои — Е. Г.*]:

> «[жесты, включающие в себя конфигурацию g r a p p o l o ‘связка, кластер’] *основываются на мотиве* ***вычленения чего-либо важного и привлечения к этому важному внимания слушающего*** <…>, ***выражают идею ‘суть’, ‘сущность’, ‘ядро’, ‘сердцевина’*** *чего-либо.* <…>
>
> *С нашей точки зрения, все конфигурации руки, включающие в себя к о л ь ц о, восходят к ситуации удержания чего-либо между кончиками большого и указательного пальцев* <…> *Мы согласны с Десмондом Моррисом* <…> *в том, что можно выделить группу жестов, которые используют конфигурацию «кольцо» и которые восходят к ‘точному захвату’* [precision grip]*, когда большой и указательный пальцы используются, чтобы взять и удержать что-либо маленькое. Как мы можем видеть,* <…> *все жесты, основанные на этой конфигурации, включают в себя* ***идею точности, четкости, ясного изложения какого-либо факта или идеи***».

Аналогичная идея (захвата, собирания) лежит и в основе группы жестов, включающих в себя конфигурацию *щ е п о т ь*. Однако, как пишет далее А. Кендон,

> «*способ захвата, а следовательно, и те семантические импликатуры, которые лежат в основе этих двух групп жестов, различны.*
> *В случае щепоти либо объект сжимается пальцами, либо пальцами*

---

[2] База данных включает в себя более 500 жестоупотреблений. В базу вошли в основном эпизоды из передачи «Гордон» (НТВ, 2002–2003 гг.), в которых в беседе с журналистом ученые обсуждали самые разные проблемы современной науки. Кроме того, материалом для статьи послужили видеозаписи, сделанные на конференции «Диалог».

поднимают объект, имеющий средний или нерелевантный размер, либо пальцами собирают в одну кучку множество маленьких объектов, каким-то образом распределенных в пространстве <…>. В случае берущего движения, характерного для к о л ь ц а, <…> поднимается вверх один очень маленький объект, или один определенный объект выбирается из группы однородных объектов».

Таким образом, основные значения по крайней мере двух конфигураций рук, связанных с соединением пальцев, достаточно хорошо и подробно описаны в литературе. Обратим, однако, внимание на следующие обстоятельства.

Во-первых, группы жестов, основанные на конфигурациях *щепоть* и *кольцо*, анализировались в основном на романском, прежде всего, на итальянском материале. Заранее не понятно, в какой степени этот анализ актуален для русской жестикуляционной системы.

Во-вторых, как показывает материал, конфигурация *соединение пальцев* чрезвычайно характерна для говорящих на русском языке, но при этом очень плохо осознается самими носителями[3].

В-третьих, исчерпывающий анализ А. Кендона, с нашей точки зрения, имеет один методологический недостаток: жест исследуется как единство 1) конфигурации руки, 2) ориентации ладони, 3) направления движения руки и 4) его кратности. С нашей точки зрения, все перечисленные параметры являются самостоятельными и могут свободно комбинироваться, причем каждый параметр передает свой компонент совокупного значения целого жеста — прагматический, референциальный, грамматический, эвиденциальный, синтаксический или лексический (см. об этом также [Крейдлин 2007: 323]). Как следствие, в анализе А. Кендона мы не можем быть уверены, что приписанные жестам *щепоть* и *кольцо* значения в конкретных разобранных примерах возникают именно благодаря данным конфигурациям руки, а не благодаря ориентации ладони, кратности жеста и расположению его в жестикуляционном пространстве.

Таким образом, в данной статье мы хотели бы понять, как устроено семейство жестов, основанных на соединении пальцев, в русской жестикуляционной системе, ввести эти жесты в широкий лингвистический оборот в качестве специальных семантических маркеров, а также попытаться понять, какие именно значения передает в данном семействе жестов именно конфигурация руки, вне зависимости от остальных жестикуляционных параметров (направление движения, ориентация в пространстве, кратность). Кроме того, у нас есть ряд уточнений, касающихся этимонов жеста, т. е. касающихся тех дожестовых движений, которые мотивируют дальнейшее метафорическое развитие жестикуляционной семантики (подробнее о жестикуляционных этимонах см. [Гришина 2014а]).

---

[3] Автор статьи лично слышал от двоих в высшей степени профессиональных лингвистов, что они никогда не используют ни жест кольцо, ни жест щепоть, причем один из собеседников сопровождал это утверждение жестом щепоть.

2. Основные конфигурации рук, включающие в себя соединение пальцев

2.1. Кольцо

На рис. 1 изображена конфигурация *кольцо*. Для нее характерно касание кончиков большого и указательного пальцев и отогнутые от ладони остальные пальцы.

Рис. 1. Кольцо

У данной конфигурации имеется два значимых компонента — пальцы, которые формируют более или менее точную окружность, и соприкасающиеся кончики большого и указательного пальца. Именно компонент *окружность* дал название всей конфигурации (*кольцо*). Парадоксальным образом, однако, именно этот компонент в жестикуляции используется очень редко — наша база данных включает 570 контекстов, и только в двух в основе семантики жеста лежит именно этот значимый компонент, иконически изображающий круглый объект или передающий идею окружности: кольцо{*детектором размером с монету*}[4]; *которые всасывают в себя* кольцо{*окружающее вещество*}.

В остальных примерах, где используется конфигурация *кольцо*, значимым компонентом являются соприкасающиеся кончики пальцев. В этом случае крайние фаланги указательного и большого пальцев, соединяясь, формируют (по мере возможности) прямую линию, что и ведет к кольцеобразной форме жеста, т. е. кольцо в данном случае имеет вынужденный характер и является

---

[4] Далее примеры даются курсивом, **зона действия жеста** (т. е. часть фразы, соответствующая ударной фазе жеста и пост-ударному удержанию, если оно есть) берется в фигурные скобки, а тип жеста дается в верхнем регистре перед фигурными скобками. (Ударной фазой жеста (по А. Кендону) называется основной этап осуществления жеста, без которого жест не считается состоявшимся; помимо обязательной ударной части в жесте выделятся экспозиция (подготовительная часть), предударное удержание (рука удерживается на стадии экспозиции перед ударной частью), постударное удержание (рука удерживается в позиции ударной части некоторое время до ретракции) и ретракция (рука возвращается в исходное дожестовое положение; все фазы осуществления жеста, кроме ударной, являются факультативными.)

не «целью» жеста, а его побочным продуктом. А центральной частью жеста является точка, образованная точным соединением двух отрезков прямой, которые задаются фалангами соприкасающихся пальцев (см. рис. 2). Таким образом, более правильным названием для жеста такой конфигурации было бы название *точь-в-точь*, а не *кольцо*, поскольку округлая конфигурация пальцев в данном случае является незначимой, вынужденной.

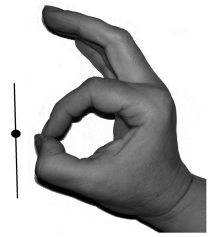


Рис. 2. Точь-в-точь

2.2. Перо (плоскогубцы и отрезок)

Конфигурация, которую мы предлагаем назвать *перо*, отличается от конфигурации *кольцо/точь-в-точь*, на первый взгляд, весьма незначительно: во-первых, остальные пальцы, кроме большого и указательного, могут быть плотно прижаты к ладони (рис. 3) или оставаться свободно, ненапряженно полусогнутыми (рис. 4); во-вторых, фигура, образованная большим и указательным пальцем, напоминает не кольцо, а перо птицы или ручки-самописки: часть фигуры, формируемая ладонью, остается округлой, а часть, формируемая подушками пальцев, может оказаться в той или иной степени вытянутой. Последнее вызвано тем, что фигура образована не с помощью кончиков пальцев, а с помощью двух подушечек (рис. 3; аналогичная фигура, напоминаем, Ж. Кальбрис была названа *finger pinch*) или с помощью ногтя большого пальца и подушечки указательного (рис. 4).

Рис. 3. *Перо*[1], или *плоскогубцы* (finger pinch по Ж. Кальбрис)

Рис. 4. *Перо*[2], или *отрезок*

В основе данной конфигурации жестов, при всей внешней похожести, лежит геометрия, отличающая их как друг от друга, так и от конфигураций *кольцо/точь-в-точь*. *Перо*[1], конфигурация, которую мы в дальнейшем будем называть *плоскогубцы*, в качестве основы для дальнейших метафорических переносов имеет идею касания, соединения двух плоскостей, задаваемых подушечками пальцев (см. рис. 5).

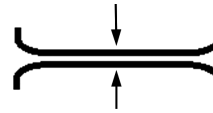


Рис. 5. *Плоскогубцы* — соединение плоскостей

Соединение двух плоскостей передает три идеи: во-первых, саму идею соединения каких-либо объектов, а во-вторых, — идею совпадения двух плоскостей в одной плоскости. В отличие от слияния двух линий в одной точке, которые лежат в основе жеста *точь-в-точь* (это слияние требует особой точности), — слияние двух плоскостей, состоящих из **множества** точек, является более простым, не требующим специального «прицеливания» действием, а посему передает не идею точного совпадения, а идею совпадения per se, вне зависимости от его точности. Третьей идеей является идея удержания пальцами какого-то маленького объекта, т. е. та идея, которая, с точки зрения А. Кендона, передается с помощью конфигурации *кольцо*, а с точки зрения Ж. Кальбрис — с помощью конфигурации *плоскогубцы* (в ее терминологии — finger pinch). Как станет ясно из последующего изложения, трактовка Ж. Кальбрис оказалась для русского материала более актуальной, чем трактовка А. Кендона.

Что касается конфигурации *перо*[2], в которой ноготь или кончик большого пальца отмечает границу небольшого участка подушечки указательного пальца, то такая конфигурация передает идею маленькой части, маленького отрезка, небольшого расстояния (в дальнейшем мы будем называть такую конфигурацию *отрезком*, см. рис. 6).

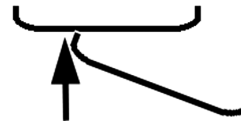


Рис. 6. *Отрезок* — маленькое расстояние

Данная конфигурация встречается и в более отчетливом, проявленном виде, хотя и достаточно редко: в этом случае кончик большого пальца отмеряет крайнюю фалангу указательного пальца практически целиком (см. рис. 7).

Рис. 7. Отрезок

На нашем материале такая конфигурация встретилась только один раз[5] и обозначала масштаб карты, т. е. размещение в одном маленьком отрезке (*один сантиметр*) большой длины (*один километр*): *то есть* $^{\text{отрезок}}${*в одном сантиметре один километр*}. В менее строгом варианте, проиллюстрированном на рис. 4, эта конфигурация, напротив, достаточно частотна.

2.3. Щепоть

Щепоть в русской жестикуляции осуществляется с помощью двух основных конфигураций — *троеперстие* (рис. 8) и *собственно щепоть* (рис. 9).

Рис. 8. Троеперстие

Рис. 9. Собственно щепоть

Различаются эти конфигурации лишь числом соединенных пальцев — три или пять.

В основе *щепоти* лежат две существенно различные идеи. Первая из них связана с тем, что плоскость задается тремя точками, а объем — количеством точек больше трех. Таким образом, три кончика пальцев в *троеперстии*

---

[5] Строго говоря, такая конфигурация встречается еще один раз, но там она является комбинацией указания указательным пальцем и конфигурации кольцо: отрезок{можешь облететь ее (гору)}: кольцо передает идею облета вокруг горы, а указательный палец чертит траекторию облета.

задают некоторую плоскость, которая является сечением имеющего определенный объем объекта. *Собственно щепоть* кончиками пальцев задает четыре или пять точек, а следовательно, и объем некоторого объекта. Таким образом, в этом случае *щепоть* может рассматриваться как минимальный по объему вариант такой конфигурации руки, как *держащая рука* (подробнее об этой конфигурации см. [Гришина 2014б]). *Держащая рука* формируется расставленными пальцами, как бы задающими объем объекта, находящего в ладони, см. рис. 10.

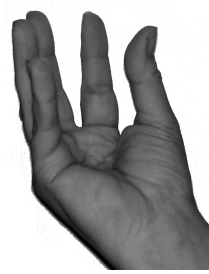


Рис. 10. Держащая рука

По мере уменьшения объекта уменьшается его объем и вес, а следовательно, в пределе объект, рассматриваемый как объемная точка, можно держать уже не в ладони, а лишь пальцами, см. на рис. 11 соответствующие трансформации от конфигурации *держащая рука* к конфигурации *собственно щепоть*. Таким образом, *щепоть* в первом варианте задает некоторый объект, который рассматривается говорящим как компактная объемная точка[6].

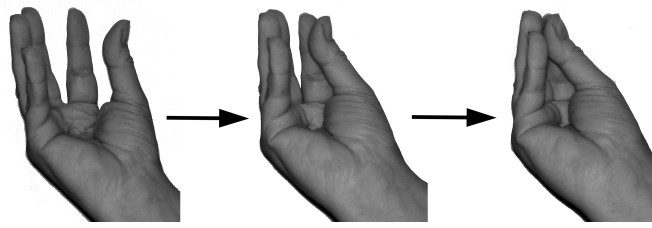


Рис. 11. Трансформация *держащая рука* → *собственно щепоть*

Вторая идея, лежащая в основе конфигурации *щепоть*, ближе всего к образу пирамиды, которую усмотрела в данной конфигурации Ж. Кальбрис (см. выше, стр. 185). Однако, с нашей точки зрения, это не просто пирамида,

[6] Надо ли говорить, что в качестве такой компактной объемной точки говорящий может рассматривать как атом, так и галактику, — к объективному размеру объекта это отношения не имеет, имеет значение лишь объем с прагматической точки зрения.

а пирамида, сложенная из векторов, направленных центростремительно к вершине пирамиды, см. рис. 12. Векторы, естественно, задаются сложенными пальцами.

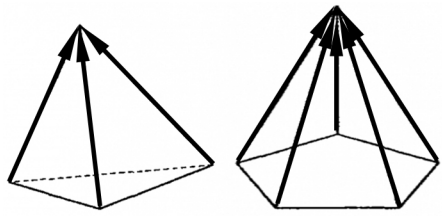


Рис. 12. Центростремительная пирамида

Такая трактовка конфигурации *щепоть* подразумевает идею центра, к которому стремится какой-то процесс, или идею базы, исходной точки, корня, из которого позже начнет развиваться некоторый процесс.

Наличие векторов в данной конфигурации и ее пирамидальное, стреловидное строение подтверждается тем, что в автодейксисе (т. е. в автоуказании на говорящего), по нашим данным, не используются конфигурации *кольцо/точь-в-точь*, однако автодейксис часто и легко осуществляется с помощью конфигурации *щепоть*. К сожалению, данных для статистических выкладок недостаточно, поскольку часто камера поставлена так, что точно рассмотреть конфигурацию руки не удается, а следовательно, не удается точно различить *щепоть*, с одной стороны, и *перо*, — с другой, однако отчетливо заметно, что ни в одном из случаев автодейксиса не используется конфигурация *точь-в-точь*, для которой характерно полное отсутствие векторной компоненты.

3. Типы контекстов, сопровождаемых соединением пальцев

3.1. Группа ‘точность’

Далее в этом разделе будут приведены типичные контексты, которые сопровождаются жестом, включающим в себя в той или иной форме соприкосновение пальцев. Примеры оформлены так, как указано в сноске 4.

Здесь следует сказать о том, что в неподготовленной речи жестикуляция очень часто осуществляется раньше, чем в речи появляется соответствующая этой жестикуляции лексема, граммема или синтаксическая конструкция (см. об этом, в частности, [Calbris 2011:54–55, 59–62 и далее; *перевод мой* — Е. Г.]: *мы видим, что информация, которая передается жестами, не является лишь дубликатом информации, выраженной с помощью слов: говорящий, используя*

все доступные ему каналы передачи смысла, дополняет одно другим <…> жест может как дополнять, так и предвосхищать речь, поскольку жест сам по себе способен передавать информацию (с. 68–69)). В результате, как можно видеть ниже, жест весьма часто осуществляется раньше, чем говорящий произносит соответствующий этому жесту речевой фрагмент.

Кроме того, следует учитывать такое явление, как catchment (подхват), т. е. временное прерывание жеста и его возвращение. В результате довольно широкого использования подхватов, в частности, для синтаксического членения устной речи (см. [Николаева 2013], соответствующая жестикуляция может прерываться на части сопровождающей жест фразы, а затем возвращаться.

И наконец, жестикуляция может несколько запаздывать относительно соответствующей лексемы или конструкции, что бывает связано 1) с тяготением ударной фазы жеста к имеющейся во фразе эмфазе, 2) с тяготением ударной фазы жеста не к конкретной лексеме, а к лексеме вместе со связанными с ней и непосредственно прилегающими к ней синтаксическими компонентами, 3) с необходимостью сделать в определенной точке фразы другой жест (т. е. ситуация, когда два жеста не могут быть исполнены одновременно в одной и той же точке высказывания, а говорящий испытывает потребность в обоих жестах).

Перечисленные выше факторы приводят к тому, что смысловой компонент, вызвавший данный жест, часто приходится искать в ближайшем контексте, а в ряде случаев жест, который выглядит как отдельное жестикуляционное событие, является всего лишь результатом подхвата предшествующего жеста и не должен анализироваться как самостоятельный (см. также далее, Раздел 4).

3.1.1. Числительные[7]

Стандартные контексты: точь-в-точь*{тоже за ноль секунд};* точь-в-точь*{это уже 1000 лет для солнца выработать такую энергию}*

Контексты с приблизительными числами[8]: плотность больше плотности воды точь-в-точь*{на 14-15 порядков}; где-то вот* точь-в-точь*{в семидесятых годах}*

Контексты с местоимениями-числительными и другими лексическими единицами, имеющими количественный компонент в значении: вот такую перо*{**величину**}; там* перо*{закачано бесконечно **много** информации};* перо*{здесь остается **масса** вопросов}*

---

[7] Отметим здесь, что в связи с причинами, которые будут описаны нами позже, в Разделе 4, мы не разделяем далее разные виды конфигурации перо (на плоскогубцы и отрезок).

[8] Приблизительность обычно передается колебательными движениями ладони, руки или перебором пальцев (см. [Гришина 2014в]), однако в ряде случаев, как мы видим, говорящий пренебрегает семой приблизительности и сопровождает жестом сему числа.

3.1.2. ‘Точность’

Далее в цитатах полужирным выделены лексемы, передающие идею точности: вам важно чтобы увидеть перо*{все **детали**}; *перо*{а доклад был сделан **как раз вот** на заре этой микроэлектроники}; *точь-в-точь*{математически} **корректно***

Компонент ‘точность’ присутствует также в конструкции уточнения формулировки (подразумевается что-то типа точнее говоря): перо*{яркий не всплеск, а яркое послесвечение после всплеска}*

3.1.3. ‘Точный выбор’

Сема ‘точный выбор’ характеризуется тем, что некоторый объект выбирается из множества однородных объектов, причем предполагается не просто выбор объекта из множества, но и противопоставление выбранного объекта не выбранным. Для воплощения идеи точного выбора используются лексические, грамматические, синтаксические средства.

Лексические средства

лексема только (только X, а не Y, Z…): собраны точь-в-точь*{**только**} из правых сахаров; государство может* собственно щепоть*{не **только** перераспределять}*

лексема именно (именно X, а не Y, Z…): и получить собственно щепоть*{**именно** то, что} ты хочешь; лексема именно может подразумеваться, но реально отсутствовать в высказывании: можно употребить в контекстах* точь-в-точь*{какого [именно] типа};* точь-в-точь*{когда [именно] существовал} их праязык*

лексемы с семантическим компонентом ‘выбрать’, т. е. найти среди множества сходных объектов нужный объект, при этом не случайно, а приложив к этому усилия: точь-в-точь*{чтобы **открыть** какое-то какое-то новое явление как правило требуются дорогие прецизионные установки}; и ему* собственно щепоть*{удалось **обнаружить**}; позволяют* троеперстие*{**выбирать** маршрут}*

лексемы с семантическим компонентном ‘особенный’, т. е. отличающийся чем-либо специфическим от однородных объектов: это было собственно щепоть*{**особенно** болезненно}; то есть вы специфичность функций* перо*{заменяете **специфичным** состоянием среды}*

Грамматические средства

превосходная степень и сходные конструкции, выделяющие один объект среди множества однородных объектов: то перо*{здесь на этой картинке он был бы ярчайшей звездой}; исследования* перо*{по **самым высокоскоростным** ударам};* перо*{одно из **блестящих** решений НАСА}*

Синтаксические средства

сложноподчиненные предложения с соотносительным словом (обычно тот) в главном предложении — эти конструкции можно соотнести с описанной выше конструкцией с лексемой именно, поскольку во многих сложноподчиненных предложениях такого типа контрольная лексема именно подставляется в главное предложение без изменения его смысла; конструкция же в целом подразумевает выделение некоего объекта с помощью соотносительного слова и описание его специфики и отличия от других однородных объектов с помощью следующего придаточного: что они распределены в плоскости этой галактики, точь-в-точь*{[именно]* ***там*** *где они и рождаются}; до* точь-в-точь*{****того*** *как} простерлось небо; или наоборот видят в данных* точь-в-точь*{****то*** *чего в них нет}.*

3.2. Группа ‘соединение’

В данную группу входят контексты, которые содержат идею соединения двух сущностей или объектов. Соединение может подразумевать физический контакт, но также и абстрактное слияние. Сюда же относится семантика симметрии или равновесия, а также контексты, которые описывают сравнение двух сущностей. Сопоставление, сравнение, равенство, симметрия, равновесие предполагают одновременную актуальность двух объектов, которая и позволяет выявить их сходство, симметричность посредством их мысленного или физического совмещения. Кроме того, в эту группу входят контексты с идеей дублирования, удвоения объекта или сущности, поскольку и в этом случае мы получаем два объекта, которые могут быть без остатка совмещены друг с другом.

3.2.1. ‘Соединение’

Семантический компонент ‘соединение’: и они точь-в-точь*{так} раз! — и* ***соединили*** *меня с этим маленьким городом; если* собственно щепоть*{****соединение*** *кристаллизуется}; и чтобы они* перо*{меж собой не* ***слипались****}*

Семантический компонент ‘связь’: мы сейчас знаем перо*{все* ***взаимодействия****}; чтобы установить* собственно щепоть*{****связь****}; они же* ***привязаны*** перо*{к определенным} событиям*

3.2.2. ‘Сопоставление’

Семантический компонент ‘сравнение’: на основе перо*{****сравнения****} некоторых таких черт; если он может все время* перо*{****сравнивать*** *и включать}; сравнительная степень: ну то есть* перо*{****чаще****} оно сопровождает сему я*

Семантический компонент ‘равенство’: перо*{что такое межъязыковая* ***эквивалентность****};* перо*{****эквивалентность****} в переводе;* собственно щепоть*{должно быть* ***уравновешено*** *правдой}*

Семантический компонент ‘дублирование’: которые перо*{могли бы **реплицироваться**}*

3.3. Группа ‘маленький объект’

Данная группа делится на три. В первую, основную группу контекстов, включены контексты, в которых теми или иными лексическими и словообразовательными средствами передана идея малости объекта. Во вторую группу вошли контексты, в которых упоминаются те или иные относительно тонкие (т. е. имеющие маленькое поперечное сечение) линии. Несколько сложнее третья группа — ‘отклонения’. Туда включены контексты, описывающие ситуацию очень маленького, по сравнению с базовой линией, отклонения от этой базовой линии.

3.3.1. ‘Маленький объект’

Лексические средства: эти часы и дни сжимаются для нас перо*{в **секунды**}; в котором* перо*{он потом отложит свои **личинки**}; и* собственно щепоть*{в каждой} **мелочи***

вытянутый объект, имеющий очень маленькое сечение: множество троеперстие*{**тонких** эффектов}; и* перо*{очень **узкая**} линия; довольно* перо*{хитрую и **тонкую** вещь};*

полное отрицание: мы не обнаружили перо*{**ни одного**} белка;* собственно щепоть*{**ни один** линейный} процесс*

Словообразовательные средства: перо*{кров**иночк**ой капель**к**ой одной}; чере**почк**и и черные и* перо*{цветные}; потому что* перо*{пыл**инк**а} на зеркале*

3.3.2. ‘Линия’

Лексические средства: перо, колебательное дугообразные движение обеими руками изнутри наружу*{около **горизонта** событий возникает};* точь-в-точь, дуга правой рукой слева направо*{любые события можно связать ломаной **линией** нулевой длины}*

3.3.3. ‘Отклонение’

Лексические средства: перо*{плотность энергии **флуктуации** вакуума};* точь-в-точь*{а атом это только какие-то небольшие **отклонения**};* перо*{**амплитуда** колебаний} 10 в минус семнадцатой сантиметра*

3.4. Группа ‘центр’

Если пирамиды, изображенные на рис. 12, сжать в плоскость вдоль вертикальной оси, то мы увидим, что обе они подразумевают понятие физического центра, который задают стремящиеся к нему векторы, см. рис. 13.

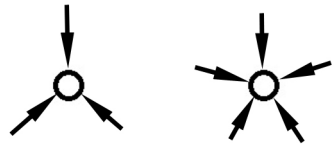


Рис. 13. Вершина пирамиды как центр

На этой геометрии жестов с компонентом *соединение пальцев* основаны следующие подгруппы значений.

Физический центр
точка, к которой направлено некоторое движение: свет упадет собственно щепоть*{**внутрь**}; оно бы* собственно щепоть*{**сколлапсировало**};*

точка, из которой направлено некоторое движение: посмотреть собственно щепоть*{есть ли там **источник** с точностью} несколько угловых секунд;* троеперстие*{**источник** который затухал}; *троеперстие*{то есть из **центральной** части керна}; потому что* собственно щепоть*{**изнутри**} свет наружу не поступает*

просто центральная зона некоторого объекта: 15 миллиардов лет назад это было все … собственно щепоть*{**в одной точке**}; что* троеперстие*{есть **база**}, есть какое-то базовое наименование*

Понятийный центр
что собственно щепоть*{**самое главное**};* собственно щепоть*{еще тут **важный** такой **момент**}; противоречат* собственно щепоть*{**сути** своей науки}; в нем* собственно щепоть*{по его **существу**, по его свойствам}*

Временной центр
собственно щепоть*{человек существовал} **до всего**;* собственно щепоть*{и оказывается вот эта вот **первичная конструкция**}; вы не разделите в этом* собственно щепоть*{**первоначале**}*

3.5. Группа ‘объект’

В этой группе с помощью соединения пальцев может обозначаться некоторый объект.

Само слово «объект»,
а также иные контексты, которые включают в себя лексемы, обозначающие некоторые материальные объекты: точь-в-точь*{вот представим **лампочку**}; должны были приносить какие-то* собственно щепоть*{археологические} **объекты***

Контексты идентификации

К группе ‘объект’ относятся также контексты, в которых некоторый объект вычленяется из окружающего мира посредством отнесения его к какому-либо классу объектов (=к множеству объектов, обладающих определенными свойствами) — но без противопоставления множеству однородных объектов, которое характерно для группы ‘точный выбор’, см. выше.

Собственно идентификация: троеперстие*{это типичные коллокации};* перо*{это эффект Допплера}; то есть это* перо*{вектор} бесконечномерного пространства;*

присвоение имени: то что мы называем перо*{переходом к непреодолимой сложности};* перо*{это вообще общая болезнь, на научном языке это называется…};*

дефиниции: а карта это собственно щепоть*{тот же снимок}; человек это божественное* собственно щепоть*{проявление с индивидуальной свободной волей};*

приписывание свойств: перо*{репликация это есть свойство исключительно биологических систем};* собственно щепоть*{у меня сын оператор};*

в качестве стандартного способа приписать свойство объекту используется сложноподчиненное предложение с определительным придаточным: для египтятина собственно щепоть*{который умел во всем…}; какой-нибудь* собственно щепоть*{возникает эффект который на самом деле трансформируется};*

к этому же типу относятся сложноподчиненные предложения, в которых предикатом в главном предложении являются ментальные глаголы, глаголы восприятия и речи, а придаточное раскрывает содержание мысли, чувства и речи, т. е. определяемым объектом здесь является факт мысли, чувства, речи, а его свойства раскрываются в придаточных предложениях: перо*{и мы* ***видим****, что там происходит}; и мы не* ***знаем*** собственно щепоть*{на сегодня какова наша вселенная}*

сюда же можно отнести контексты, которые можно назвать активированная анафора, — они включают в себя конструкции вот + анафорическое местоимение + предикация; группа вот + анафорическое местоимение формирует некий объект из предшествующего текста, а предикация присваивает этому объекту некоторое свойство, т. е. относит его к некоторому классу: перо*{****вот там*** *они и могут излучить эти гамма-кванты};* собственно щепоть*{то* ***вот этот*** *раздражитель приобретал};* собственно щепоть*{что* ***вот этот*** *степенной закон} представляет собой*

4. Жесты и типы контекстов: статистические распределения

В этом разделе мы приведем данные о том, как распределяются жесты, основанные на соединении пальцев, между разными группами контекстов. Здесь нужно сделать несколько пояснительных замечаний.

Во-первых, если еще раз вернуться к рис. 1, 3–4, 7–9, то можно заметить, что различия между разными конфигурациями жестов весьма незначительны:

- жест *кольцо/точь-в-точь* при редуцированном исполнении весьма слабо отличается от жеста *перо* со свободными, не прижатыми к ладони пальцами — отличие касается только степени напряженности отогнутых от ладони пальцев и близости фигуры, сложенной из указательного и большого пальцев, к кольцу.
- жест *плоскогубцы* и жест *отрезок* при редуцированном исполнении отличаются всего лишь расположением подушечек большого и указательного пальца друг относительно друга
- жест *троеперстие* иногда очень трудно отличить от жеста *перо*, поскольку при мгновенном соприкосновении пальцев часто неясно, большой палец касается только указательного, или одновременно указательного и среднего
- точно так же при быстром и редуцированном исполнении большой палец соприкасается с указательным и средним, что вроде бы формирует конфигурацию *троеперстие*, но при этом вполне может быть, что это результат редукции жеста *собственно щепоть*, а безымянный палец и мизинец просто не успели подтянуться для осуществления жеста в полном объеме.

Во-вторых, как мы уже писали выше, жесты в спонтанной неподготовленной речи осуществляются не в точечном режиме, когда жест выровнен точь-в-точь с отдельным словом, а в режиме 1) предвохищения речи, 2) затягивания (пост-ударное удержание), 3) запаздывания, когда жест осуществляется уже после вызвавшего его семантического компонента (см. выше, стр. 193). Таким образом, жест в спонтанной речи может осциллировать вокруг вызвавшего его семантического компонента и может также захватывать множество «посторонних» семантических компонентов.

И в-третьих, зона действия жеста может содержать разнонаправленные семантические компоненты. Например, во фрагменте *пришли к размерам меньше микрона* компонент *микрон* относится к группе ‘числительное’ и одновременно — к группе ‘маленький объект’; к последней группе относится также компонент *меньше*; фрагмент *вот такую величину* отсылает одновременно к конструкции активированная анафора (группа ‘объект’) и к группе ‘числительное’, контекст *а доклад был сделан как раз вот на заре этой микроэлектроники* может отсылать к группе ‘точность’ и к группе ‘маленький объект’. И так далее, таких примеров очень много.

Учитывая перечисленные три фактора, мы можем утверждать, что закономерности в соотношении между жестом и контекстом могут иметь **только статистический характер**, и эта неопределенность непреодолима по самому объективному положению вещей. Именно поэтому при анализе жестикуляционной семантики анализ отдельных примеров может быть весьма поучительным

и чрезвычайно много объясняющим, но уверенность, что исследователю удалось определить основные семантические компоненты жеста, может дать только статистически достоверный размер базы данных.

Имея в виду сказанное выше, приведем в табл. 1 данные о том, как распределяются жесты, включающие в себя соединение пальцев, между названными выше семантическими группами[9]. В таблице полужирным выделены ячейки, в которых реальные данные существенно отличаются от ожидаемых в среднем, если бы параметры не были связаны, в **большую** сторону, *курсивом* — в меньшую. В скобках приводятся значения χ^2 для данной ячейки.

Таблица 1

Тип жеста / **Тип контекста**	**кольцо/ точь-в-точь**	**перо**	**троеперстие**	**собственно щепоть**
1. активированная анафора	2	*5 (χ^2=1,35)*	3	**7 (χ^2=1,9)**
2. ‘объект’	1	*6 (χ^2=1,2)*	**4 (χ^2=1,54)**	**8 (χ^2=2,36)**
3. ментальные и прочие глаголы + придаточное или именная группа	*0 (χ^2=5,31)*	18	4	**14 (χ^2=3)**
4. ‘идентификация’	*4 (χ^2=3,14)*	30	**11 (χ^2=1,87)**	19
5. ‘центр’	*4 (χ^2=1,05)*	*7 (χ^2=10,36)*	**12 (χ^2=9,14)**	**22 (χ^2=10,8)**
6. ‘линия’	3	**25 (χ=4,73)**	2	*3 (χ^2=3,23)*
7. ‘маленький объект’	12	**56 (χ^2=3,1)**	9	*13 (χ^2=3,78)*
8. ‘отклонение’	1	**15 (χ^2=3,41)**	1	*2 (χ^2=1,53)*
9. ‘соединение’	3	**15 (χ^2=1,2)**	2	*3 (χ^2=1,25)*
10. ‘сопоставление’	1	**17 (χ^2=5,2)**	1	*1 (χ^2=3,12)*
11. ‘точность’	**12 (χ^2=9,7)**	*12 (χ^2=1,34)*	2	8
12. ‘точный выбор’	**10 (χ^2=1,7)**	18	4	13
13. числительные	**22 (χ^2=17,34)**	26	*3 (χ^2=2,44)*	12
χ^2=129,41, $p \leq 2{,}65^{-12}$, распределения достоверны, параметры связаны				

Мы видим, что статистические пики в табл. 1 образовали три группы (1–5, 6–10, 11–13). Сгруппируем эти строки в соответствии с типами контекстов, перечисленными в предшествующем разделе. Одновременно обратим внимание на то, что столбцы «троеперстие» и «собственно щепоть» ни в одной строке не противоречат друг другу, т. е. либо пик соответствует пику (строки 2, 5), либо пик соответствует статистически незначимым данным (строки 1, 3, 4); точно так же и падения в этих столбцах либо соответствуют друг другу, либо падение соответствует статистически незначимым данным; это позволяет нам

[9] Безусловно, уверенности, что нам удалось определить все типы использования жестов, у нас нет никакой. Не исключено, что при увеличении базы в два или более раз, 1) появятся новые группы, которые не удалось определить уверенно на нашей базе данных (особенно это касается синтаксических конструкций), 2) по-новому перераспределятся старые данные. Однако объем базы в 500 с лишним примеров дает нам надежду, что основные типы использования данного набора жестов нами «ухвачены».

склеить два этих столбца[10]. В результате произведенных трансформаций получаем табл. 2, где выявленные статистически закономерности, на наш взгляд, более наглядны.

Таблица 2

Тип жеста **Группа контекстов**	**кольцо/** **точь-в-точь**	**перо**	**щепоть**
‘Точность’	**44 (χ^2=25,31)**	*56 (χ^2=2,76)*	*42 (χ^2=1,64)*
‘Соединение’	4	**32 (χ^2=5,55)**	*7 (χ^2=4,65)*
‘Маленький объект’	16	**96 (χ^2=9,76)**	*30 (χ^2=8,74)*
‘Центр’	4	*7 (χ^2=10,34)*	**34 (χ^2=19,52)**
‘Объект’	*7 (χ^2=8,52)*	59	**70 (χ^2=9)**
χ^2=109,86, p≤4,06^{-20}, распределения достоверны, параметры связаны			

Итак, мы видим следующее.

1. Жест *точь-в-точь*, в полном соответствии со своей геометрией и внутренней формой, используется для сопровождения контекстов из группы ‘точность’ (числительные, собственно ‘точность’ и ‘точный выбор’).
2. Конфигурация *перо* (напомним, сюда включены две разные конфигурации — *плоскогубцы* и *отрезок*) имеет два пика на двух разных типах контекстов. В группе контекстов ‘соединение’, по нашему мнению, основной является конфигурация *плоскогубцы* (рис. 3, 5), в которой актуальным является сам факт соприкосновения подушечек пальцев. Что касается группы ‘маленький объект’, то здесь актуальны обе конфигурации — и *отрезок* (рис. 4, 6, в основе лежит образ небольшого расстояния, отмеренного ногтем или кончиком большого пальца на линии, которая задана подушечкой указательного пальца), и *плоскогубцы* (в основе лежит образ щипцов, пинцета, зажавших какой-то маленький объект).
3. И, наконец, конфигурация *щепоть* также имеет два пика — на группах ‘объект’ и ‘центр’. Представляется, что для группы ‘объект’ актуальна *щепоть* как предел сжатия конфигурации *держащая рука* (см. рис. 11), т. е. объект — это нечто, что держат сжатые пальцы. Для группы же ‘центр’ актуальна *щепоть* как образ пирамиды, у которой векторы ребер сходятся в центре-вершине (см. рис. 12).

Таким образом, соприкосновение пальцев в русской жестикуляции передает пять основных значений — ‘точность’ (жест *точь-в-точь*), ‘соединение/сопоставление’ (жест *плоскогубцы*), ‘маленький объект’ (жесты *плоскогубцы* и *отрезок*), ‘объект’ и ‘центр’ (жест *щепоть*).

---

[10] Склеиванию столбцов точь-в-точь и перо мешает строка 11 (‘точность’), где пик для точь-в-точь соответствует падению для перо.

В заключение отметим следующий немаловажный момент. Является более или менее общепринятой идея, что этимоном большинства, если не всех речевых (co-speech) жестов является повседневная деятельность человека, его дожестовые, чисто функциональные действия (например, взять маленький объект для жеста *кольцо*, выбрать один предмет из множества однородных для жеста *grappolo*, и так далее). Это в значительной степени верно, спорить можно лишь о том, правильно ли было определен исследователем соответствующий этимон. Однако анализ жестов показывает, что на жестикуляционную метафорику в очень значительной степени влияет эвклидова геометрия, которая сама базируется на повседневном практическом опыте человека, однако представляет собой следующий, гораздо больший уровень абстракции.

Литература

1. *Calbris G.* (2011). Elements of meaning in gesture. Benjamins,Amsterdam/Philadelphia, 2011
2. *De Jorio A.* (1832/2000). Gesture in Naples and Gesture in Classical Antiquity. A translation of La mimica degli antichi investigate nel gestire napoletano (1832), and with Introductione and Notes, by Adam Kendon. Bloomington: Indiana Univ. Press, 2000
3. *Diadori P.* (1990). Without Words. 100 Italian Gestures [Senza Parole. 100 Gesti degli italiani]. Rome: Bonacci, 1990
4. *Efron D.* (1982). Gesture, Race and Culture. The Hague: Mouton and Co., 1972
5. *Kendon A.* (1995). Gestures as illocutionary and discourse structure markers in southern Italian conversation // Journal of Pragmatics, 23: 247–279
6. *Kendon A.* (2004). Gesture. Visible Action as Utterance. Cambridge Univ. Press, 2004
7. *Morris D.* et al. (1979). Gestures: Their Origins and Distribution. London: Jonathan Cape, 1979
8. *Munari B.* (1963). Add-ins to Italian Dictionary [Supplemento al dizionario italiano]. Milan: Muggiani, 1963
9. *Гришина Е. А.* (2014а). Вертикальная ось в жестикуляции: лингвистический аспект // Русский язык в научном освещении, № 27 (в печати)
10. *Гришина Е. А.* (2014б). Жесты и прагматические характеристики высказывания // Мультимодальные коммуникации: теоретические и эмпирические исследования. М., 2014, с. 25–47
11. *Гришина Е. А.* (2014в). Круги и колебания: семантика сложных траекторий в русской жестикуляции // Когнитивные подходы к языку: Сборник статей. М., 2014 (в печати)
12. *Крейдлин Г. Е.* (2007). Механизмы взаимодействия вербальных и невербальных единиц в диалоге II а. Дейктические жесты и их типы // Труды международной конференции «Диалог 2007»: компьютерная лингвистика и интеллектуальные технологии. М., 2007, с. 320-327.
13. *Николаева Ю. В.* (2013). Иллюстративные жесты в русском дискурсе. Дисс. на соискание степени кандидата филол. наук. М., МГУ, 2013

References

1. *Calbris G.* (2011). Elements of meaning in gesture. Benjamins,Amsterdam/Philadelphia, 2011
2. *De Jorio A.* (1832/2000). Gesture in Naples and Gesture in Classical Antiquity. A translation of La mimica degli antichi investigate nel gestire napoletano (1832), and with Introductione and Notes, by Adam Kendon. Bloomington: Indiana Univ. Press, 2000
3. *Diadori P.* (1990). Without Words. 100 Italian Gestures [Senza Parole. 100 Gesti degli italiani]. Rome: Bonacci, 1990
4. *Efron D.* (1972). Gesture, Race and Culture. The Hague: Mouton and Co., 1972
5. *Grishina E. A.* (2014a). Vertical axis in Gesticulation from the Linguistic Point of View [Vertikal'naja os' v zhestikuljatsii: lingvisticheskij aspekt] // Russkij jazyk v nauchnom osveshchenii, № 27 (forthcoming)
6. *Grishina E. A.* (2014б). Gesticulation and Pragmatics [Zhesty i pragmaticheskie harakteristiki vyskazyvanija] // Mul'timodal'nye kommunikatsii: teoreticheskie i empiricheskie issledovanija. M., 2014, p. 25–47
7. *Grishina E. A.* (2014в). Circles and Swings: Semantics of Composite Paths in Russian Gesticulation [Krugi i kolebanija: semantika slozhnyh traektorij v russkoj zhestikuljatsii] // Kognitivnye podhody k jazyku: Sbornik statej. M., 2014 (forthcoming)
8. *Kendon A.* (1995). Gestures as illocutionary and discourse structure markers in southern Italian conversation // Journal of Pragmatics, 23: 247–279
9. *Kendon A.* (2004). Gesture. Visible Action as Utterance. Cambridge Univ. Press, 2004
10. *Krejdlin G. E.* (2007). Mechanisms of interation between verbal and nonverbal units in a dialog II a. Deictic gestures and their types [Mehanizmy vzaimodejstvija verbal'nyh i neverbal'nyh edinits v dialoge II a. Dejkticheskie zhesty i ih tipy // Trudy vezhdunarodnoj konferentsii "Dialog 2007": komp'juternaja lingvistika i intellektual'nye tehologii. M., 2007, p. 320-327.
11. *Morris D.* et al. (1979). Gestures: Their Origins and Distribution. London: Jonathan Cape, 1979
12. *Munari B.* (1963). Add-ins to Italian Dictionary [Supplemento al dizionario italiano]. Milan: Muggiani, 1963
13. *Nikolaeva Ju. V.* (2013). Illustrative gestures in Russian discourse [Illjustrativnye zhesty v russkom diskurse]. Diss. na soiskanie stepeni kandidata filol. nauk. M., MGU, 2013

К СОЗДАНИЮ ЧАСТОТНОГО СЛОВАРЯ ЗНАЧЕНИЙ СЛОВ[1]

Иомдин Б. Л. (iomdin@ruslang.ru)

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

Лопухина А. А. (nastya-merk@yandex.ru)

Институт русского языка
им. В. В. Виноградова РАН, Москва, Россия

Носырев Г. В. (grigorij-nosyrev@yandex.ru)

Яндекс, Москва, Россия

В докладе на материале русских существительных с предметными значениями обосновывается необходимость создания частотного словаря значений слов. Предлагаются методы приближенного определения частот, основанные на анализе данных опросов информантов и аннотировании наиболее частотных коллокаций в большом корпусе текстов (в настоящей работе был использован самый объемный на сегодняшний день корпус RuTenTen11, интегрированный в систему Sketch Engine). Такой словарь мог бы быть востребован в различных компьютерно-лингвистических приложениях (в частности, для вероятностного разрешения многозначности в отсутствие контекста), при создании обучающих ресурсов, в традиционной толковой лексикографии. Исследования наборов значений многозначных слов и их сравнительной частотности представляют и теоретический интерес для изучения эволюции лексической системы языка.

Ключевые слова: семантика, лексикография, многозначность, полисемия, омонимия, предметная лексика, опросы, эксперименты, частота, статистические методы

[1] Работа выполнена при частичной финансовой поддержке Программы фундаментальных исследований отделения историко-филологических наук РАН «Язык и литература в контексте культурной динамики», гранта РГНФ №13-04-00307а и гранта НШ-3899.2014.6 для поддержки научных исследований, проводимых ведущими научными школами РФ.

TOWARDS A WORD SENSE FREQUENCY DICTIONARY

Iomdin B. L. (iomdin@ruslang.ru)
Vinogradov Russian Language Institute
of the Russian Academy of Sciences, Moscow, Russia;
National Research University
Higher School of Economics, Moscow, Russia

Lopukhina A. A. (nastya-merk@yandex.ru)
Vinogradov Russian Language Institute
of the Russian Academy of Sciences, Moscow, Russia

Nosyrev G. V. (grigorij-nosyrev@yandex.ru)
Yandex, Moscow, Russia

Analyzing several Russian nouns denoting everyday life objects, we explain why a word sense frequency dictionary is necessary. Techniques of calculating the approximate frequencies are proposed, based on the analysis of native speaker surveys and the annotation of the most frequent collocations in a large text corpus (we used the huge RuTenTen11 corpus integrated into the Sketch Engine system). A word sense dictionary could be used in a variety of NLP tasks, in particular for a probabilistic word sense disambiguation without available context, in creating second language learning resources, as well as in academic lexicography. Besides, studies of sense sets of polysemous words and their comparative frequencies are important for the linguistic theory, because they shed light on the evolution of the lexical system.

Key words: semantics, lexicography, ambiguity, polysemy, homonymy, everyday life vocabulary, surveys, experiments, frequency, statistical techniques

1. Введение

В работе, опубликованной в материалах предыдущей конференции «Диалог», мы отмечали: «в последнее время большое значение, в частности, в компьютерной лингвистике, придается созданию частотных словарей и списков слов. К сожалению, большинство имеющихся частотных списков составляются из вокабул, но не отдельных лексем (то есть слов, взятых в определенном значении). Между тем очевидно, что разные лексемы одной и той же вокабулы частотны в очень разной степени; столь же очевидно, что составление частотных списков лексем представляет собой существенно более трудную и практически не автоматизируемую задачу» [Иомдин и др. 2013: 319].

Проблема отсутствия частотных словарей значений признается не только в лингвистике, но и, например, в сфере педагогики, в частности, при составлении списков слов для изучения в разных классах школ; ср. “Some problems are inevitable when word frequency is the primary source for identifying words. <...> A word that has different meanings is listed only once. For example, whether *bank* means financial institution, edge of a river, or angle of an airplane is not taken into account. *B-a-n-k* appears one time on the list, and its associated frequency represents all the different meanings. In other words, there is no way to get the frequency of the word *bank* meaning a financial institution” [Beck at al. 2013: 21]. Действительно, в существующих частотных словарях обычно не различаются не только разные лексемы, но и омонимы (ср., в частности, [Ляшевская и Шаров 2009]). Нет информации о частотности значений и в переводных словарях, что затрудняет работу с ними (особенно в случае сильно многозначных слов, когда недостаточно знакомый с языком читатель вынужден строить много равновероятных гипотез о значении незнакомого слова). Востребованность такой информации кажется бесспорной. Симптоматично, что система статистического машинного перевода Google Translate с 31.10.2012 стала визуализировать частотность различных вариантов перевода (но не значений переводимого слова!), хотя пока выдаваемые результаты достаточно спорны, во всяком случае с точки зрения адекватности переводных эквивалентов (ср. рис. 1–2):

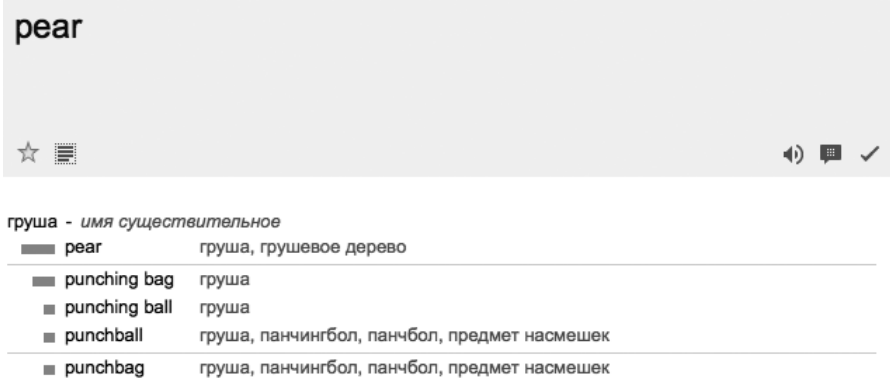


Рис. 1. Варианты перевода слова *груша* в Google Translate

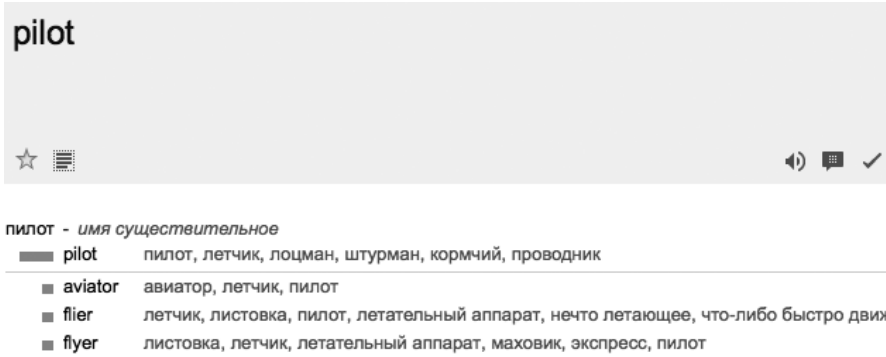


Рис. 2. Варианты перевода слова *пилот* в Google Translate

В настоящей работе делается попытка обосновать необходимость создания частотных словарей значений слов и предлагаются первые подходы к достижению этой цели.

2. Материал

Хотя задача создания частотных словарей значений предполагает саму возможность разделения слова на значения, на самом деле это сопряжено с большими трудностями. Известно, что носители языка часто дают не совпадающие ответы о наборах значений одних и тех же слов, разные лексикографы принимают разные решения на одном и том же материале, а в теоретической семантике высказываются мнения и о принципиальной недискретности полисемии (ср. обсуждение этих проблем, в частности, в работах [Апресян 1974/1995; Kilgarriff 1993; Pustejovsky 1996; Wilks 1998; Lin, Ahrens 2005; Зализняк Анна 2006; Иомдин 2014]). В практической лексикографии, однако, тем не менее делить слова на значения приходится. В частности, это очень существенно для переводных словарей, где объективным критерием выделения значений может считаться различие переводных эквивалентов. Кроме того, неопределенность полисемии неодинакова для разных частей речи и для разных семантических полей. Так, например, наборы значений и структура многозначности предлогов (*в*, *по* и др.) или глаголов положения в пространстве (*лежать*, *висеть* и др.) существенно различаются в разных словарях и для разных носителей, в то время как в предметной лексике значения часто выделяются более четко, поскольку возможно явное указание референтов, имеющих различные свойства (ср. *визитка* ‘предмет одежды’, ‘сумочка’, ‘визитная карточка’, *банан* ‘фрукт’, ‘предмет одежды’, ‘водный аттракцион’, ‘оценка’).

Для настоящей работы в качестве пробного материала мы отобрали семь достаточно частотных многозначных вокабул, имеющих конкретные предметные значения и описанных в Активном словаре русского языка (АС) [Апресян и др. 2014]: *альбом, билет, блок, вешалка, вилка, винт, горшок*, поставив

задачу выявить наиболее частотное значение каждого из этих слов. Для этого мы а) проанализировали существующие толковые словари, б) провели опрос информантов и в) исследовали наиболее частотные словосочетания с указанными словами на корпусном материале.

3. Исследование

3.1. Анализ толковых словарей

Для того чтобы определить набор и порядок значений исследуемых слов с точки зрения академической лексикографии, мы проанализировали их описание в восьми толковых словарях[2]. Каждое значение мы называем при помощи короткого инварианта, составленного по материалам толкований из АС, а для тех значений, которые в АС отсутствуют, приводим упрощенное толкование из соответствующего словаря.

Для слова ***альбом*** в СУш, БАС 1, МАС и БТС приводятся два значения: 1) ‘тетрадь или книга с чистыми листами’ и 2) ‘книга с репродукциями или фотографиями’. В СОШ, БАС 2, СЕф и АС, помимо этих двух, появляется третье значение — ‘собрание музыкальных произведений’ (в СЕф оно описывается в двух разных толкованиях: ‘комплект из нескольких пластинок’ и ‘собрание песен, записанное на магнитную ленту’).

Более интересной оказывается эволюция значений слова ***билет***: в СУш приводится всего одно значение — ‘документ, удостоверяющий право на услугу’; это же значение в БАС 1 идет под номером 3), а перед ним оказываются 1) ‘документ, удостоверяющий членство в организации’ и 2) ‘денежный знак’. У данного слова в БАС 1 есть также значения 4) ‘жребий, набор заданий на экзамене’ и 5) ‘документ, удостоверяющий личность’. В словарях, составленных позже (МАС, СОШ, БТС), значений всего четыре и они располагаются в следующем порядке: 1) ‘документ, удостоверяющий право на услугу’, 2) ‘документ, удостоверяющий членство в организации’, 3) ‘денежный знак’, 4) ‘набор заданий на экзамене’. В АС значения 3) и 4) поменяны местами, т. е. более редким считается значение ‘денежный знак’. В БАС 2 кроме этих четырех значений сохраняется 5) ‘документ, удостоверяющий личность’ с уточнением «в дореволюционной России».

Почти во всех словарях для слова ***блок*** выделяются два омонима: моносемичный со значением ‘устройство для подъема грузов’ и полисемичный, порядок и набор значений у которого варьируется от словаря к словарю. Первым значением второго омонима в СУш, БАС, МАС, СОШ, БТС и БАС 2 является значение ‘группа организаций’, в АС же данное значение — третье. Первыми, т. е. самыми актуальными, значениями в АС оказываются значения 1.1) ‘большой

---

[2] Приведем список словарей в хронологическом порядке: СУш —1935–1940 гг., БАС —1948–1965 гг., МАС — 1981–1984 гг., СОШ — 1992 г., БТС — 1998 г., БАС 2–2004–2006 гг., СЕф — 2006 г. и АС — 2014 г.

сплошной брусок или плита' и 1.2) 'строительный элемент', которые, видимо, появились у слова ***блок*** достаточно давно (есть в МАС). Более новые значения — 2.1) 'несколько связанных помещений', 2.2) 'часть устройства' и 2.3) 'часть компьютерной программы' — либо зафиксированы в словарях, изданных позднее (значение 2.1 впервые встречается в БТС), либо пока не отмечены другими словарями (например, значение 2.3). В БТС в качестве значения 2) дается 'совокупность однородных предметов, понятий, явлений и т. п.', а в АС это большое значение разделено на два: 'совокупность однородных физических объектов' (*блок сигарет*) и 'совокупность информационных объектов' (*блок статей*) и стоит в слове ***блок*** последним. В БТС и БАС 2, кроме двух описанных омонимов, выделяется третий — в БТС он представлен одним значением 'защитный прием', а в БАС 2 к данному значению добавляется 'то же, что блокпост'.

Слово ***вешалка*** оказывается хорошим примером того, как особое употребление лексемы может развиться в самостоятельное значение. Так, в СУш у слова два значения: 'приспособление с крючками' и 'петелька на одежде', они есть и во всех остальных рассмотренных словарях. В БАС у первого значения отмечено употребление 'планка с крючком, плечики', которое уже в СОШ дается как самостоятельное второе значение. В АС значения слова ***вешалка*** поданы в следующем порядке: 1.1) 'приспособление с крючками', 1.2) 'помещение гардероба', 2) 'планка с крючком, плечики' (с образным употреблением 'высокая худая женщина'), 3) 'петелька на одежде'.

Для слова ***вилка*** все словари выделяют как наиболее актуальное значение 'столовый прибор'. Вторым в СУш, БАС 1, МАС, СОШ, БТС, БАС 2 и СЕф дается значение 'приспособление или устройство с раздвоенным концом'; в АС этот смысл выражен в значении 2) 'штепсель'. Более редкие значения — 'серия попыток' и 'положение в шахматной игре' (зафиксированы уже в БАС 1). Кроме того, достаточно новым можно считать значение 'крайние значения параметра'.

Первые два значения слова ***винт*** — 'предмет с резьбой для соединения деталей' и 'вращающаяся деталь' — описаны практически одинаково и в СУш, и в АС. Для данного слова изменения проявляются в развитии новых значений: 'элемент фигурного катания' и 'завинчивающаяся жестяная пробка' (БТС), 'спортивное упражнение, состоящее из вращения вокруг вертикальной оси тела' (БАС 2), 'жесткий магнитный диск' (СЕф), 'наркотик' (АС) — а также в утрате неактуальных значений ('винтовка' — в СУш и 'карточная игра' — отдельное значение в СУш и БАС 1, омоним в МАС и БАС 2, в АС нет).

У слова ***горшок*** во всех словарях выделяются смыслы 'сосуд для приготовления пищи', 'сосуд для комнатных растений' и 'сосуд для естественных отправлений', причем в большей части словарей порядок именно такой. Различие состоит в статусе этих смыслов: в СУш, БАС 1 и МАС все это — употребления внутри одного значения, в БАС 2 последний смысл выделяется в отдельное значение, а в СЕф и АС все представлены в виде разных значений. В СОШ у слова ***горшок*** два значения — 1) 'сосуд для приготовления пищи' и 2) 'сосуд для естественных отправлений' — а словосочетание *цветочный горшок* подается как фразема. Кроме этого, в БАС 2 есть устаревшее значение 'сосуд с горючими веществами, употреблявшийся в военном деле'.

Таким образом за 80 лет, прошедшие со времени выхода первого из рассмотренных толковых словарей, у всех семи слов в них менялись либо порядок значений, либо набор значений, либо статус приписываемых им смыслов (от употребления к значению или наоборот). В меньшей степени изменения коснулись слов ***альбом*** и ***горшок***; больше всего изменений отражено в словарных статьях ***блок*** и ***билет***.

3.2. Эксперимент с информантами

В феврале 2014 года мы провели эксперимент в интернете, состоящий из двух частей. Первая часть эксперимента выглядела следующим образом: информанты-носители русского языка должны были описать самое частое, по их мнению, значение каждого из данных выше семи слов и затем привести другие их значения (в том порядке, в котором они приходят в голову, без долгих размышлений). Значения разрешалось описывать разными способами: давать определение (можно краткое или неточное) или приводить примеры. При выполнении эксперимента не допускалось использование словарей и других источников. По завершении первой части информанту выдавалась вторая часть, где приводились краткие описания значений тех же семи слов, в основном адаптированные из АС. Для каждого значения каждого слова необходимо было оценить, часто ли оно, по мнению информанта, встречается в современных текстах на русском языке («часто», «средне», «редко», «никогда или почти никогда»).

В эксперименте приняло участие более 700 человек из более чем 100 населенных пунктов (около 30 % мужчин и около 70 % женщин), их средний возраст составил 33 года. 53 % информантов указали, что имеют законченное высшее образование, еще 16 % — степень кандидата наук и 2 % — степень доктора наук; среди остальных информантов было 17 % студентов, 7 % аспирантов, 3 % школьников и 2 % людей с законченным средним образованием.

Приведем некоторые результаты эксперимента.

В Таблице 1 в среднем столбце приведены данные о значениях, которые приводили информанты в первой части опроса, когда список значений еще не был им предъявлен («укажите самое частое, по вашему мнению, значение»), в правом столбце — процент информантов, оценивших соответствующее значение из выданного набора как «частое» во второй части опроса («часто ли это значение, по вашему мнению, встречается в современных текстах на русском языке»).

Таблица 1. Результаты опроса информантов

	привели первым	**оценили как «частое»**
Альбом		
Книга с репродукциями или фотографиями	42 %	67 %
Тетрадь или книга с чистыми листами	41 %	56 %
Собрание музыкальных произведений	8 %	77 %

	привели первым	оценили как «частое»
Билет		
Документ, удостоверяющий право на услугу	96%	99%
Набор заданий на экзамене	3%	61%
Документ, удостоверяющий членство где-л.	0%	39%
Денежный знак	0%	2%
Блок		
Строительный элемент	21%	44%
Упаковка сигарет	9%	54%
Устройство для подъема грузов	4%	16%
Часть устройства	3%	43%
Группа организаций	3%	38%
Спортивный прием	2%	—[3]
Вешалка		
Приспособление для хранения одежды	90%	84%
Помещение гардероба	<1%	21%
Худая женщина	<1%	11%
Неприятная ситуация	<1%	—[4]
Вилка		
Столовый прибор	95%	99%
Штепсель	2%	81%
Крайние значения параметра	0%	13%
Положение в шахматной игре	0%	11%
Серия попыток	0%	1%
Винт		
Предмет с резьбой для соединения деталей	68%	90%
Вращающаяся деталь	20%	60%
Наркотик	<1%	16%
Жесткий диск	<1%	—[5]
Горшок		
Сосуд для приготовления пищи	41%	39%
Сосуд для комнатных растений	16%	89%
Сосуд для естественных отправлений	16%	56%

[3] Это редкое значение, приведенное информантами, отсутствует в АС, поэтому не предлагалось в эксперименте изначально: ср. *Я поставил блок рукой, а левой стукнул ему в скулу* (С. Шаргунов).

[4] Это новое сленговое значение, приведенное информантами, отсутствует в АС; ср. *С ребенком там совсем вешалки.. ни в магазин, ни в аптеку…* (Интернет-форум).

[5] Этот омоним, использующийся в компьютерном сленге (сокращение от *винчестер*), отсутствует в АС; ср. *Т. е. для того чтобы записать по сетке видео на винт, подключенный к плееру, достаточно пары нажатий на пульте* (Компьютерра.Лаб, 02.02.2010).

Как видно, значение, которое информанты выделяют как наиболее частотное до предъявления им полного списка значений, не всегда получает максимальную оценку при оценивании всего списка. Скажем, для слова *альбом* значение ‘собрание музыкальных произведений’ в первой части опроса назвали самым частотным лишь около 8% информантов, однако во второй части опроса именно это значение чаще всего оценили как «часто используемое» (77%); для слова *горшок* значение ‘сосуд для комнатных растений’ в первой части опроса назвали самым частотным лишь меньше 16% информантов, однако во второй части опроса именно это значение чаще всего оценили как «часто используемое» (89%).

Сложности предсказуемо возникли при наличии у слова большого числа трудно отделимых от друга и плохо формулируемых значений (ср. *блок*), а также в случаях, когда информанты давали слишком общее толкование, не позволявшее определить, какое именно значение имелось в виду (так, для слова *вешалка* в качестве первого значения чаще всего указывали ‘то, на что вешают одежду’, что не дает возможности различить значения ‘приспособление с крючками’ и ‘планка или каркас с крючком, плечики’).

Таким образом, однозначно определить наиболее частотные значения в представлении информантов удалось для трех слов: *билет* (‘документ, удостоверяющий право на услугу’), *вилка* (‘столовый прибор’) и *винт* (‘предмет с резьбой для соединения деталей’).

3.3. Эксперимент с корпусом

Помимо опросов носителей, в современной лексикографической работе необходимо широкое использование корпусных данных. Самым объемным на сегодняшний день корпусом русского языка является собранный из текстов интернета корпус RuTenTen11, насчитывающий около 20 млрд словоупотреблений. Неоценимым удобством этого корпуса для лексикографов является его интеграция в систему Sketch Engine (www.sketchengine.co.uk, Kilgarriff et al. 2004), позволяющую получать списки наиболее частотных коллокаций с данным словом. Мы исследовали такие списки для указанных семи слов (первые десять самых частотных коллокаций в каждой синтаксической конструкции, предлагаемой системой Sketch Engine, в случае, если для этой конструкции система выдавала больше десяти коллокаций; в противном случае конструкция не учитывалась). Каждой коллокации два лексикографа-соавтора статьи независимо друг от друга сопоставили одно из значений, содержащихся в АС (случаи, где сделать это было трудно или невозможно, отмечались особо)[6]. Приведем пример такого аннотирования по одной конструкции для одного слова (всего для

[6] Известно, что эксперты-носители языка по-разному оценивают конкретные употребления, что затрудняет работу над созданием аннотированных корпусов текстов (inter-judge variance). Так, в работе [Fellbaum et al. 1997] было показано, что на решения аннотаторов влияло расположение значений в списках, которые им выдавались; в работе [Snyder and Palmer 2004] несогласие аннотаторов связывается со слишком дробным делением на значения, принятым в тезаурусе WordNet [Иомдин 2014].

каждого слова использовались данные от 10 до 30 конструкций, то есть аннотировались от 100 до 300 частотных словосочетаний с каждым словом).

Таблица 2. Набор значений для слова вешалка (адаптированный из AC)

1	Приспособление с крючками (*У двери стоит вешалка*)
2	Помещение гардероба (*Театр начинается с вешалки*)
3	Планка или каркас с крючком, «плечики» (*Платье на вешалке*)
4	Худая женщина
5	Петелька на одежде (*пришить вешалку к куртке*)

Таблица 3. Пример работы аннотаторов

Самые частотные коллокации в конструкции с прямым дополнением:

	Значение (аннотатор 1)	**Значение** (аннотатор 2)
прибивать	1	1
пришивать	5	5
смастерить	1	1
опрокинуть	1	1
уронить	1,3	1,3
повесить	1,3	1,3
закрепить	1	1
спланировать	1	2
крепить	1	1
прикреплять	1	1

Затем для каждого значения было подсчитано суммарное количество отнесенных к нему коллокаций, помноженное на их общую частотность в корпусе. Если какая-то коллокация была отнесена к двум или нескольким значениям, ее частота учитывалась при подсчете частотности каждого из этих значений. Если результаты двух аннотаторов различались, подсчитывалось среднее арифметическое между получающимися частотами. См. формулы:

$$f_s(i) = \frac{1}{N_a} \sum_a \sum_w f(w, c = i_a)$$

$$f_S = \sum_w f_S(w)$$

$$F = \sum_S F_s$$

$$f(i) = \sum_S \left(\frac{f_S(i)}{f_S} \cdot \frac{F_S}{F} \right)$$

N_a — Количество участвующих в разметке аннотаторов
$f_S(w)$ — Частота коллокации *w* в синтаксической конструкции *s*
$f_S(i)$ — Усредненная по аннотаторам частота класса *i* для синтаксической конструкции *s*
f_S — Суммарная частота всех рассматриваемых коллокаций для данной синтаксической конструкции
F_S — Частота синтаксической конструкции в корпусе
F — Суммарная частота всех рассматриваемых синтаксических конструкций *s*

Таким образом были получены следующие результаты (число в процентах показывает, какая доля исследованных словосочетаний предположительно относится к соответствующему значению):

Таблица 4. Результаты аннотирования коллокаций

	F_S
Альбом	
Собрание музыкальных произведений	80%
Книга с репродукциями или фотографиями	38%
Тетрадь или книга с чистыми листами	5%
Билет	
Документ, удостоверяющий право на услугу	95%
Документ, удостоверяющий членство где-л.	4%
Набор заданий на экзамене	3%
Денежный знак	<1%
Блок[7]	
Часть устройства	41%
Устройство для подъема грузов	23%
Строительный элемент	22%
Большой сплошной брусок или плита	21%
Часть компьютерной программы	18%
Несколько связанных помещений	12%
Совокупность текстов	10%

	F_S
Группа организаций	9%
Упаковка сигарет	3%
Вешалка	
Приспособление с крючками	82%
Планка с крючком, «плечики»	39%
Помещение гардероба	14%
Петелька на одежде	2%
Вилка	
Столовый прибор	53%
Штепсель	25%
Крайние значения параметра	<1%
Положение в шахматной игре	<1%
Серия попыток	<1%
Винт	
Предмет с резьбой для соединения деталей	57%
Вращающаяся деталь	41%
Предмет в форме спирали	5%
Наркотик	<1%
Горшок	
Сосуд для комнатных растений	74%
Сосуд для приготовления пищи	34%
Сосуд для естественных отправлений	30%

[7] Мы не приводим здесь данных об омониме *Блок* (фамилия), коллокации с которым также встречались среди частотных (*подражать Блоку, Блок и Маяковский* и т. п.)

Как видим, этот эксперимент позволяет достаточно надежно выявить наиболее частотные и наименее частотные значения для каждого из семи слов (если считать, что исследованный корпус адекватно отражает их распределение в языке)[8].

4. Заключение

Итак, задача создания частотного словаря значений слов, во всяком случае в области предметной лексики, кажется теоретически выполнимой. Интересно, что если эксперимент, в котором информанты пытаются определить наиболее частотное значение спонтанно, далеко не всегда дает показательные результаты, то оценки частотности, данные носителями языка при предъявлении словарного списка значений, достаточно хорошо соотносятся с результатами экспертной оценки коллокаций, полученных в корпусе. Как кажется, сочетание этих методик может быть взято за основу при работе над частотным словарем значений. Первым шагом мог бы стать словарь, в котором для каждого многозначного слова было бы указано наиболее частотное значение, если оно надежно выделяется на основе экспериментов с носителями и корпусных исследований.

С практической точки зрения, такой словарь мог бы быть использован в различных компьютерно-лингвистических приложениях (в частности, для вероятностного разрешения многозначности в отсутствие контекста, см. также об этой проблеме работу [Иомдин 2014]), в создании лексических минимумов, разговорников, учебников и обучающих ресурсов, в традиционной толковой лексикографии (в частности, описанные методы уже используются при работе над Словарем бытовой терминологии (см. [Иомдин 2011, Иомдин и др. 2012, 2013]), в котором дается информация о частотности).

С теоретической точки зрения, регулярные исследования наборов значений многозначных слов и их сравнительной частотности представляют интерес для изучения эволюции лексической системы языка. Так, интересно, что сразу в нескольких рассмотренных словах на первый план выходят значения, утратившие связь с этимологически первоначальным смыслом [по Фасмер 1986]. Ср. *альбом*, восходящее к лат. album ‘открытый лист для сбора подписей’, букв. ‘белый’: в наиболее частотном сейчас «музыкальном» значении идея чистоты утрачивается, а разрабатывается идея совокупности; *блок*, восходящее к германскому корню **blok-* ‘твердый ствол’: в наиболее частотных значениях утрачивается идея твердости, но разрабатывается идея ‘часть целого’; *горшок*, производное от *горн* ‘плавильная печь’: в наиболее частотном сейчас значении

[8] Для более объективных результатов при работе над словарем необходимо учитывать данные разных источников, что пока осложняется отсутствием подобной Sketch Engine системы автоматического подбора частотных коллокаций, во всяком случае для русского языка. Кроме того, при наличии соответствующей разметки корпуса возможны и интересны также исследования особенностей социального, регионального, возрастного, гендерного распределения частотности значений.

‘сосуд для комнатных растений’ утрачивается связь с печью. Динамика частоты значений важна и для сравнительных исследований, прежде всего на материале региональных вариантов, диалектов и родственных языков.

Литература

1. *Апресян* 1974/1995 — Апресян Ю. Д. Лексическая семантика. Синонимические средства языка. М., 1974 (2-е изд.: М., 1995).
2. *Апресян* и др. 2014 (АС) — Активный словарь современного русского языка. А-Г / Под ред. Ю. Д. Апресяна. М.: «Языки славянских культур», 2014 (в печати).
3. БАС *1* — Словарь современного русского литературного языка: В 17-ти т. / Изд-во АН СССР; Под ред. В. И. Чернышева. М., Л., 1948–1965.
4. БАС *2* — Большой академический словарь русского языка / ИЛИ РАН; Под ред. К. С. Горбачевича. М, СПб.: «Наука», 2004–2006. Тт. 1–4.
5. БТС — Большой толковый словарь русского языка / Сост., гл. ред. С. А. Кузнецов. СПб.: Норинт, 1998.
6. *Зализняк* 2006 — Зализняк Анна А. Многозначность в языке и способы ее представления. М., 2006.
7. *Иомдин* 2011 — Иомдин Б. Л. Материалы к словарю-тезаурусу бытовой терминологии. СВИТЕР: образец словарной статьи // Слово и язык. Сборник статей к восьмидесятилетию академика Ю. Д. Апресяна. Отв. ред. И. М. Богуславский, Л. Л. Иомдин, Л. П. Крысин. М.: «Языки славянских культур», 2011. С. 392–406.
8. *Иомдин* и др. 2012 — Иомдин Б. Л., Лопухина А. А., Пиперски А. Ч., Киселева М. Ф., Носырев Г. В., Рикитянский А. М., Васильев П. К., Кадыкова А. Г., Матиссен-Рожкова В. И. Словарь бытовой терминологии: новые проблемы и новые методы // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2012» (Бекасово, 30 мая — 3 июня 2012 г.). Вып. 11 (18). М.: РГГУ, 2012. С. 213–226.
9. *Иомдин* и др. 2013 — Иомдин Б. Л., Лопухина А. А., Панина М. Ф., Носырев Г. В., Вилл М. В., Зайдельман Л. Я., Матиссен-Рожкова В. И., Винокуров Ф. Г., Выборнова А. Н. Маг вел мот: изменения в языке на материале бытовой терминологии // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2013» (Бекасово, 29 мая — 2 июня 2013 г.). Вып. 12 (19): в 2 т. Т. 1. М.: РГГУ, 2013, с. 311–324.
10. *Иомдин* 2014 — Многозначные слова в контексте и вне контекста // Вопросы языкознания, 2014, №4 (в печати).
11. *Ляшевская и Шаров* 2009 — О. Н. Ляшевская, С. А. Шаров, Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.
12. МАС — Словарь русского языка: В 4-х т. /АН СССР, Ин-т рус. яз.; Под ред. А. П. Евгеньевой. М.: Русский язык, 1985–1988.

13. *СЕф* — Ефремова Т. Ф. Большой современный толковый словарь русского языка. В 3-х т. М., АСТ, Астрель, 2006.
14. *СОШ* — Ожегов С. И. и Н. Ю. Шведова. Толковый словарь русского языка. М.:Азъ, 1992.
15. *СУш* — Толковый словарь русского языка / Под ред. Д. Н. Ушакова. М.: Гос. ин-т Сов.энцикл.; ОГИЗ; Гос. изд-во иностр. и нац. словарей, 1934–1940.
16. *Фасмер* 1986 — Фасмер М. Этимологический словарь русского языка. Тома I–IV. М., 1986.
17. *Beck* at al. 2013 — Beck, Isabel L., Margaret G. McKeown, and Linda Kucan. Bringing words to life: Robust vocabulary instruction.Guilford Press, 2013.
18. *Fellbaum* et al. 1997 — Fellbaum C., Grabowski K., Shari L. Analysis of a Hand-Tagging Task. In: Proceedings of ANLP-97 Workshop on Tagging Text with Lexical Semantics. 1997.
19. *Kilgarriff* 1993 — Kilgarriff A. Dictionary word sense distinctions: an enquiry into their nature // Computers and the humanities. 1993. V. 26. No. 1–2.
20. *Kilgarriff* et al. 2004 — Kilgarriff A., Rychly, P., Smrz, P., & Tugwell, D. (2004). ITRI-04-08 The Sketch Engine. Information Technology, 105, 116.
21. *Lin, Ahrens* 2005 —Lin C. C., Ahrens K. How many meanings does a word have? Meaning estimation in Chinese and English // J. W. Minett, W.S.-Y. Wang (eds). Language acquisition, change and emergence: Essays in evolutionary linguistics. Hong Kong, 2005.
22. *Pustejovsky* 1996 — Pustejovsky J. Lexical semantics: The problem of polysemy. Oxford, 1996.
23. *Snyder and Palmer* 2004 — Snyder B., Palmer M. The English all-words task // Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text. 2004. P. 41–43.
24. *Wilks* 1998 — Wilks Y. Senses and texts // Computational linguistics and Chinese language processing. 1998. V. 3. No. 2.

References

1. *Apresjan Ju. D.* (1974/1995), Lexical semantics. The synonymical means of language [Leksicheskaja semantika. Sinonimicheskie sredstva jazyka]. Moscow.
2. *Apresjan Ju. D.* (ed.). (2014), Active Dictionary of Modern Russian. A-G [Aktivnyj slovar‘ sovremennogo russkogo jazyka. A-G]. Jazyki slavjanskih kul’tur, Moscow.
3. *Chernyshov V. I.* (ed.). (1948–1965), Large Contemporary Dictionary of Standard Russian [Slovar’ sovremennogo russkogo literaturnogo jazyka]. Moscow, St. Petersburg.
4. *Gorbachevich K. S.* (ed.). (2004–2006), Large Academic Dictionary of Russian [Bol‘shoj akademicheskij slovar’ russkogo jazyka]. Moscow, St. Petersburg.
5. *Efremova T. F.* (2006), Large Contemporary Explanatory Dictionary of Russian [Bol‘shoj sovremennyj tolkovyj slovar’ russkogo jazyka]. AST, Astrel’, Moscow.
6. *Evgenyeva A. P.* (ed.). (1981–1984), Russian Language Dictionary [Slovar’ russkogo jazyka]. Russkij jazyk, Moscow.

7. *Iomdin B. L.* (2011), Materials for the thesaurus of Russian everyday life terminology. SWEATER: a sample dictionary entry [Materialy k slovarju-tezaurusu bytovoj terminologii. SVITER: obrazets slovarnoj stat'i]. Slovo i jazyk. Sbornik statej k vos'midesiatiletiju akademika Ju.D. Apresjana [The word and the language. A collection of papers to commemorate Academician Apresjan's 80th annivesrary]. Jazyki slavjanskih kul'tur, Moscow, pp. 392–406.
8. *Iomdin B. L., Lopuhina A. A, Piperski A. Ch., Kiselëva M. F., Nosyrev G. V., Rikitjanskij A. M., Vasil'jev P.K., Kadykova A. G., Matissen-Rozhkova V. I.* (2012), Thesaurus of Russian everyday life terminology: new problems and new techniques [Slovar' bytovoj terminologii: novye problemy i novye metody]. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012". Bekasovo, pp. 213–226.
9. *Iomdin B. L., Lopuhina A. A, Panina M. F., Nosyrev G. V., Vill M. V., Zajdel'man L. Ja., Vinokurov F. G., Matissen-Rozhkova V. I., Vybornova A. N.* (2013), Mag vel mot: language innovations in everyday life terminology [Mag vel mot: izmenenija v jazyke na materiale bytovoj terminologii]. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2013". Bekasovo, pp. 311–324.
10. *Iomdin B. L.* (2014), Polysemous words within and without context [Mnogoznachnyje slova v kontekste i vne konteksta]. Voprosy jazykoznanija [Issues in Linguistics]. Vol. 4. Moscow (in print).
11. *Kilgarriff A.* (1993), Dictionary word sense distinctions: an enquiry into their nature. In: Computers and the humanities. V. 26. No. 1–2.
12. *Lin C. C., Ahrens K.* (2005). How many meanings does a word have? Meaning estimation in Chinese and English. In: J. W. Minett, W. S.-Y. Wang (eds). Language acquisition, change and emergence: Essays in evolutionary linguistics. Hong Kong.
13. *Lyasevskaya O. N., Sharov S. A.* (2009) A Frequency Dictionary of Modern Russian (based on National Russian Corpus) [Chastotnyj slovar' sovremennogo russkogo jazyka]. Moscow: Azbukovnik.
14. *Kuznetsov S. A.* (ed.). (1998), Large Explanatory Dictionary of Russian [Bol'shoj tolkovyj slovar' russkogo jazyka]. Norint, St. Petersburg.
15. *Ozhegov S. I., Shvedova N. Yu.* (1992), Explanatory Dictionary of Russian [Tolkovyj slovar' russkogo jazyka]. Az, Moscow.
16. *Pustejovsky J.* (1996), Lexical semantics: The problem of polysemy. Oxford.
17. *Ushakov D. N.* (ed.). (1934–1940), Explanatory Dictionary of Russian [Tolkovyj slovar' russkogo jazyka]. OGIZ, Moscow.
18. *Vasmer M.* (1986), Etymological dictionary of Russian. I-IV [Etimologicheskij slovar' russkogo jazyka]. Moscow.
19. *Wilks Y.* (1998), Senses and texts. In: Computational linguistics and Chinese language processing. V. 3. No. 2.
20. *Zalizniak Anna* (2006), Polysemy in language and its representation [Mnogoznachnost' v jazyke i sposoby ee predstavlenija]. Moscow.

ВАЛЕНТНОСТИ РУССКИХ ПРЕДИКАТНЫХ СУЩЕСТВИТЕЛЬНЫХ И МИКРОСИНТАКСИЧЕСКИЕ КОНСТРУКЦИИ[1]

Иомдин Л. Л. (iomdin@iitp.ru)

Институт проблем передачи информации
им. А. А. Харкевича РАН, Москва, Россия

Иомдин Б. Л. (iomdin@ruslang.ru)

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

В статье рассматриваются такие способы реализации валентностей русских предикатных существительных в некоторых типах конструкций (в основном в бытийных конструкциях типа *Мне нет необходимости сдавать экзамен*), при которых эти предикатные существительные синтаксически прямо не связаны со словами, реализующими их валентности. В этих случаях недостаточно описать модель управления существительного, чтобы обеспечить корректную семантическую интерпретацию конструкции (при анализе) или адекватное заполнение валентностей (при синтезе). Для каждого такого слова полная информация о заполнении валентностей в рамках соответствующей конструкции должна помещаться в словаре.

Ключевые слова: валентная структура предикатных слов, микросинтаксис, нетривиальная реализация валентностей

[1] Данная работа была частично поддержана грантами РФФИ №12-07-00663 и №13-06-00756, РГНФ №13-04-00307а и программой фундаментальных исследований Президиума РАН «Корпусная лингвистика». Авторы выражают грантодателям искреннюю признательность.

VALENCIES OF RUSSIAN PREDICATE NOUNS AND MICROSYNTACTIC CONSTRUCTIONS

Iomdin L. L. (iomdin@iitp.ru)
Kharkevich Institute for Information Transmission Problems
of the Russian Academy of Sciences, Moscow, Russia

Iomdin B. L. (iomdin@ruslang.ru)
Vinogradov Russian Language Institute
of the Russian Academy of Sciences, Moscow, Russia;
National Research University
Higher School of Economics, Moscow, Russia

The paper discusses valency realizations of Russian predicate nouns in certain types of syntactic constructions (mainly, existential ones like *Mne net neobxodimosti sdavat ekzamen* 'There is no need for me to take the exam'; lit. 'to me there is no necessity...') where these realizations are not directly linked with the nouns concerned. In these cases, subcategorization frames of nouns are insufficient to account for the correct semantic interpretation of the construction in text analysis, or the adequate choice of valency implementation in text generation. For every word, detailed information on how its valencies are implemented within particular constructions should be supplied in the dictionary.

Key words: valency structure of predicate words, microsyntax, nontrivial valency implementation

Вводные замечания

Как известно, заполнение семантических валентностей предикатного слова в реальном тексте может значительно отличаться от канонической ситуации, когда эти валентности выражаются словами, синтаксически непосредственно зависящими от данного предикатного слова, т. е. выступают как его **активные** синтаксические валентности (Богуславский 2008, Boguslavsky 2009), отмечаемые в модели управления.

В настоящей статье, продолжающей нашу серию исследований нетривиальных языковых явлений, связанных с валентной структурой слов (Iomdin, Iomdin 2011, 2014, Иомдин, Иомдин 2013), мы рассмотрим некоторые микросинтаксические конструкции, в которых одним из элементов является предикатное слово, а другими элементами являются слова, реализующие его валентности (в простейшем случае — одну такую валентность).

Разумеется, таких конструкций очень много. Мы ограничимся, однако, одним, хотя и весьма представительным их классом: экзистенциальными, или, в терминах Н. Д. Арутюновой (1976), бытийными конструкциями, вершиной которых является глагол *быть* X (в эксплицитной или нулевой форме), возможно, с отрицанием (или в виде склеенного с отрицанием глагола *нет*), при котором в качестве субъекта выступает некоторое предикатное существительное Y и которому подчиняется по крайней мере одна группа P, выражающая валентность Y, как в примерах (1)–(9)[2]:

(1) *Вам* [P1] *нет* [X] *нужды* [Y] *искать* [P2] *правду, потому что она вам дана с самого начала* (Д. Быков).

(2) *Прощения* [Y] *мне* [P] *нет* [X] *, да я и не прошу его* (М. Шишкин).

(3) *Всему* [P] *есть* [X] *предел* [Y].

(4) *И по-прежнему всем* [P] *есть* [X] *дело* [Y] *до так и не найденных сокровищ царя Соломона* (Д. Рубина).

(5) *Радости* [P] *не было* [X] *конца* [Y].

(6) *Не было* [X] *пощады* [Y] *ни бедным земледельцам* [P]*, ни женщинам* (Н. Эйдельман).

(7) *— И ночью при луне мне* [P] *нет* [X] *покоя* [Y]*, зачем потревожили меня?* (М. Булгаков).

(8) *К вам* [P1] *у меня* [P2] *больше нет* [X] *никаких претензий* [Y].

(9) *Во мне* [P] *нет* [X] *страха* [Y].[3]

Активная реализация валентностей

Ниже нас будет интересовать следующий вопрос: какие именно валентности предикатных существительных заполняются именными и предложными группами в конструкциях указанного типа? Нетрудно убедиться в том, что определить это по модели управления предикатного существительного в общем случае невозможно.

---

[2] Здесь и далее большинство примеров заимствовано из НКРЯ (http://ruscorpora.ru), SynTagRus (http://ruscorpora.ru/search-syntax.html), а также частично из корпуса RuTenTen (https://www.sketchengine.co.uk).

[3] В принципе круг бытийных конструкций более широк — они, например, могут включать модальные слова, ср. *Ему нет прощения и Ему не может <не должно> быть прощения*; бытийные глаголы могут заменяться квазисинонимами, ср. *Ему не нашлось прощения*, и т.д. Мы, однако, будем рассматривать в первую очередь именно базовые бытийные конструкции.

Посмотрим, например, на предложения (1) и (2): в первом из них присутствует слово *нужда*, имеющее две основных валентности (экспериенцера[4] 'кто нуждается' и объекта 'в чем нуждаются'), а во втором — слово *прощение*, актантная структура которого состоит из трех валентностей (субъекта 'кто прощает', адресата 'кого прощают' и объекта 'за что прощают'). Если эти валентности реализуются активно, т. е. заполняющие их слова синтаксически подчиняются непосредственно рассматриваемому слову, то они выражаются для слова *нужда*, соответственно,

- родительным падежом (экспериенцер), ср.

(10) *...учитывая **нужду Войновича** в улучшении жилищных условий… собрание решило предоставить ему первую освободившуюся двухкомнатную квартиру* (В. Войнович)*;*

(11) *За службу твою я тебя наградила, граф. Но **нужду государства** я не забыла тоже* (Э. Радзинский)*;*

- предложной группой, вводимой предлогом *в* (12), или — в редких случаях — инфинитивом (13) (объект):

(12) *Верьте мне, милый друг, что **нужда в деньгах** есть признак неправильности жизни* (Л. Н. Толстой).

(13) *Особенно большое значение для предпринимателей имеет **нужда сохранить** коммерческую тайну* («Арбитражный и гражданский процессы», 2004.10.25)

Для слова *прощение* его основные валентности выражаются, соответственно,

- творительным (14) или родительным (15) падежами, либо предложной группой, вводимой предлогом *от* (16) (субъект); ср.

(14) ***прощение** отца **сыном** и сына отцом* (http://forum.kinopoisk.ru)

(15) *Он чувствовал, что в этой жизни он добился всего, но счастья нет, и нет именно потому, что он не смог добиться **прощения сестры**. И все его попытки за тридцать лет сблизиться с ней и получить от нее прощение пока ни к чему не приводили* (Ф. Искандер)*;*

---

[4] Метки семантических ролей валентностей предикатных слов, приводимые здесь, достаточно условны; авторы далеки от того, чтобы считать атрибуцию этих ролей окончательной. Мы используем их в первую очередь для того, чтобы разные валентности одного слова легче было различить читателю, и не стремимся в этом пункте к строгой научной чистоте. Для наших целей можно было бы использовать и другие инвентари семантических ролей; к примеру, для характеристики валентной структуры слов нужда и *прощение* можно было бы воспользоваться метками, приводимыми во FrameNet для соответствующих английских слов *need* (cognizer 'осознающий' и requirement 'требование') и *forgiveness* (judge 'судья', evaluee 'оцениваемый' и offence 'проступок').

(16) — *Не надобно мне ни чести твоей, ни* ***прощения от тебя*** (Ю. Герман)*;*

- родительным (17) или, значительно реже, дательным падежом (18) (адресат); ср.

(17) *Зачастую* ***прощение того****, кто провинился, нам намного нужнее, чем ему* (Ю. Ковалева)*;*

(18) *Но туда приехал уполномоченный от советского правительства и огласил Указ Президиума Верховного Совета:* ***прощение всем эмигрантам!*** (А. Солженицын)*;*

- предложной группой, вводимой предлогом *за* (19), опять-таки родительным (20) или — чрезвычайно редко — дательным падежом (21) (объект); ср.

(19) *Царь Небесный пошлет мне* ***прощение за прегрешенья*** (Б. Окуджава)*;*

(20) *… За то, что мне — прямая неизбежность — /* ***Прощение обид****, / За всю мою безудержную нежность, / И слишком гордый вид…* (М. Цветаева).

(21) *Ступай и матушке скажи, что я / К духовнику покаяться пошла / В том, что отца так сильно рассердила, И получить* ***прощение грехам*** (В. Шекспир, пер. Т. Щепкиной-Куперник).[5]

Реализация валентностей в бытийных конструкциях

Посмотрим теперь, каким образом валентности этих же слов выражаются в бытийных конструкциях. Уже из примеров (1) и (2) видно, что по крайней мере одна валентность слов *нужда* и *прощение* выражается здесь не так, как это имеет место при активном заполнении валентностей, а именно, дательным падежом (причем соответствующий актант синтаксически зависит не от предикатного существительного, а от бытийного глагола). При этом в (1) так выражается экспериенциальная валентность слова *нужда*, а в (2) — валентность адресата слова *прощение*.

Совершенно очевидно, что «вычислить», как именно реализуется та или иная валентность предикатного слова X в случае его появления в бытийной конструкции, по стандартной модели управления X невозможно. В частности, тот факт, например, что адресат слова *прощение* может выражаться при нем и дательным падежом, как в (18), мало нам поможет: если даже предположить,

---

[5] Разумеется, информация, необходимая для полноценного описания поведения этих слов в тексте, не исчерпывается перечислением способов реализации их валентностей. Мы, в частности, не рассматриваем здесь вопрос о надежной идентификации валентностей в случае совпадения способов их выражения (например, реализации всех трех валентностей слова *прощение* родительным падежом), об ограничениях, накладываемых на одновременное выражение валентностей одного слова теми или иными способами и др. Этот факт не влияет на дальнейшее изложение.

что тут из двух вариантов — родительного и дательного падежей — выбирается дательный просто потому, что в (2) слово *мне* синтаксически подчиняется глаголу *нет*, при котором родительный падеж уже есть (в нем стоит само слово *прощение*), то в предложении (1) дательный падеж экспериенцера *вам* заимствовать из модели управления слова *нужда* не удастся совсем: его там попросту нет.

Таким образом, валентную структуру предикатного слова, входящего в состав бытийной конструкции как субъект бытийного глагола, необходимо характеризовать отдельно, в дополнение к ее базовой валентной структуре.

Рассмотрим подробнее несколько таких конструкций. Начнем с уже обсужденных существительных *нужда* и *прощение*.

В бытийной конструкции, сформированной словом *нужда*, его валентности реализуются следующим образом:

- экспериенцер реализуется дательным падежом, как в (1), или предложной группой P1, вводимой предлогом *у*:

(22) ***У него*** [P1] *не было нужды* ***в другом человеке*** [P2] (А. Иличевский)*;*

(23) *И молчаливый посетитель улыбнулся и ему сказал: «****У меня*** [P1] *нет нужды* ***ставить*** [P2] ***тебе вопросы****, мне достаточно на тебя глядеть»* (митрополит Антоний (Блум)).

В обоих случаях именная или предложная группа, реализующая данную валентность, синтаксически подчиняется бытийному глаголу:

- объект, как и в случае активного заполнения этой валентности, реализуется предложной группой P2, вводимой предлогом *в*, управляющим предложным падежом (22), или инфинитивом P2 (23). При этом данная группа или инфинитив могут либо подчиняться слову *нужда*, как в (22) или (23), либо зависеть от бытийного глагола (24–25):

(24) *Я даже приехала сюда утром и все ходила возле твоего дома, будто у тебя* ***во мне*** *была нужда* (В. Орлов)*;*

(25) ***Защищать меня*** *не было нужды: обсуждение шло благосклонно* (Л. Чуковская).

В бытийной конструкции со словом *прощение* валентности последнего реализуются так:

- субъект реализуется предложной группой P1, вводимой предлогом *у*, как в (26)–(27), ср.

(26) *Но предательству* [P3] ***у нас*** [P1] *нет прощения, а предателям нет и не должно быть места на нашей земле* (В. Войнович).[6]

---

[6] Тот факт, что группа *у нас* в (26) может пониматься и как локатив (‘в нашей стране’, ‘в нашем обществе и т.д.), разумеется не исключает и субъектной интерпретации этой группы.

- адресат реализуется дательным падежом, как в (2), или предложной группой P2, вводимой предлогом *для*, как в

(27) *Чехов — это самое главное — окружал нежной своей поэзией первых — растерянных, ленивых, ищущих, — и не было у него* [P1] *прощения* ***для вторых*** [P2] *— нашедших, успокоившихся, уверенных* (К. Чуковский).

- объект реализуется предложной группой P3, вводимой предлогом *за* (28) или дательным падежом P3 (26), (29):

(28) ***За такое*** [P3] *никогда не будет прощения;*

(29) *Не может быть прощения* ***тому*** [P3]*, что запечатлено в оперативном приказе Ежова № 00486 от 15 августа 1937 года.* (А. Яковлев).

Внимательно проанализировав приведенные выше примеры типа (1–9), мы легко убедимся, что аналогичная картина наблюдается и в бытийных конструкциях, содержащих другие предикатные существительные. Разумеется, в деталях конструкции с конкретными словами могут различаться. В частности, существуют слова, которые в определенных значениях способны выступать исключительно в составе бытийных конструкций.

Таково, например, слово *дело* в значении ‘интерес’, где экспериенцером выступает слово P в дательном падеже, как в (30)–(32):

(30) *Никому* [P1] *здесь не было* [X1] *до него дела* [Y1] *— это оседлым людям* [P2] *есть* [X2] *дело* [Y2] *друг до друга, а странникам никогда* (Д. Быков).

(31) *Какое мне* [P] *ØБЫТЬ,наст* [X] *дело* [Y] *до всех до вас?*

(32) *Он сказал, что физике* [P] *нет дела* [Y]*, подтверждает ли она философию* (В. Гроссман).

(33) *И ей* [P] *не было* [X] *дела* [Y]*, что ноша называется душа* (Ю. Нагибин).

Характерно, что слово *дело* в этом значении обладает нетривиальными управляющими свойствами: помимо экспериенцера, оно имеет валентность стимула, реализуемую предложной группой, вводимой предлогом *до,* как в (30) или (31), либо разнообразными типами придаточных (32–33). Фиксировать эти управляющие свойства можно только с учетом функционирования слова в бытийных конструкциях. Так выглядит словарная статья соответствующей лексемы в редактируемом сейчас втором выпуске Активного словаря русского языка (тт. 3–4)[7]:

---

[7] Автор этой словарной статьи — Б. Л. Иомдин. Об Активном словаре русского языка см. Апресян и др. 2010, 2014.

дело 5.2, разг.
ПРИМЕРЫ. *Мне нет дела до ваших проблем*; *Какое вам дело до него?*

ЗНАЧЕНИЕ.'Интерес, который человек А1 проявляет к человеку или ситуации А2'.

УПРАВЛЕНИЕ.
А2 *до* РОД: *дело до детей <до их отношений>*;
ВОПР: (*какое кому-л.*) *дело, кто она такая*.

КОНСТРУКЦИИ. Употребляется исключительно в конструкциях с глаголом *быть* (чаще в отрицательных или вопросительных), при котором А1 выражается дополнением в форме ДАТ: *Отцу* (А1) *нет дела до него* (А2), *Какое тебе* (А1) *дело, кто ко мне приходит* (А2)?

Другими примерами существительных, определенные значения которых конструктивно обусловлены и ограничиваются бытийными конструкциями, являются:

а) слово *вера* 'доверие', где экспериенцер Р выражается дательным падежом (34) или предложной группой с *к* (35):

(34) *Чем быстрее сдамся — тем меньше мне будет* ***веры*** [Р]*!* (В. Кунин)*;*

(35) *Нет веры* ***к вымыслам чудесным*** [Р]*, рассудок все опустошил* (Ф. И. Тютчев)*;*

б) слово *слово* (в форме мн. ч.) 'возможность найти адекватный вербальный ответ', где субъектная валентность Р1 выражается предложной группой, вводимой предлогом *у,* а валентность стимула Р2, в частности, придаточным, вводимым союзом *чтобы* (36) или инфинитивным оборотом (37). В этой бытийной конструкции отрицание обязательно:

(36) *У меня* [Р1] *нет слов, чтобы* [Р2] *описать тот ужас, который охватил нас, когда мы услышали на поляне голоса пиратов* (В. Губарев)*;*

(37) *У нее* [Р1] *не было слов осмыслить это* [Р2]*; об этом невозможно было никому рассказать* (Д. Рубина).

В целом картина поведения бытийных конструкций, стержнем которых являются предикатные существительные, сводится к следующему. Поскольку в этих конструкциях по крайней мере некоторые актанты этих существительных синтаксически отрываются от них, естественно ожидать, что их реализации становятся менее индивидуальными и, так сказать, более «натуральными», тяготеющими к прототипическому смыслу предлога или падежа. Имеется две таких основных реализации: это 1) предложная группа типа *у Х-а* для — в широком смысле — субъектных валентностей и 2) дательный

падеж *Х-у для* — в широком смысле — адресатных или бенефактивных валентностей.[8] В качестве типичного примера оппозиции этих валентностей можно привести пару *У отца нет прощения* ‘Отец не может простить кого-л’ и *Отцу нет прощения* ‘Кто-л. не может простить отца’.

Однако реальная ситуация оказывается гораздо менее однозначной: часто однотипные и даже идентичные валентности близких по смыслу слов выражаются по-разному или допускают различную вариативность, ср.

(38) а) *У меня нет желания делать что-л.,*

при невозможности

(38) б) **Мне нет желания делать что-л.,*

и

(39) а) *У меня нет охоты делать что-л.,*

что допускает перифразу

(40) б) *Мне нет охоты делать что-л.*

Ср. также

(40) *У вас нет резона отказываться = Вам нет резона отказываться,*

(41) *У вас нет необходимости соглашаться = Вам нет необходимости соглашаться,*

но только

(42) а) *У вас нет возможности отказаться,*

но не

(42) б) **Вам нет возможности отказаться*.

Все эти факты однозначно указывают на то, что описание валентного поведения предикатных слов в бытийных конструкциях в значительной степени требует индивидуального подхода к каждому слову.

---

[8] Отметим для полноты картины существование еще нескольких нетривиальных типов реализаций валентностей предикатных слов в бытийных конструкциях — с помощью предлогов *в*, *на* и *за*; ср. *Во мне не было страха; Видит Бог, вины на мне нет*. (М. Шишкин); *На нем нет крови* (здесь слово кровь выступает в особом значении, близким к значению ‘убийство’); *За тобой есть должок* и др.

Эволюция дательного падежа субъекта

Если мы пристальнее взглянем на материал обсуждаемых здесь бытийных конструкций, особенно с точки зрения корпусных данных, то обнаружим любопытное явление: очевидное сокращение в последние десятилетия сферы употребления дательного падежа как выразителя субъектной валентности по крайней мере **некоторых** предикатных слов и замещение его предложно-падежной конструкцией типа *у*+S,род. Речь идет все о той же бытийной конструкции. В короткой статье невозможно подробно излагать доказательства этого тезиса, поэтому мы ограничимся лишь несколькими примерами. В НКРЯ присутствуют тексты XIX — начала XX века, в которых фигурирует заполнение субъектной валентности P предиката Y (слов *возможность*, *желание, выбор, силы, надежда*) дательным падежом :

(43) *Сильному артисту* [P] *есть возможность* [Y] *настроить себя на тот тон чувств и положений, которые в Лире и Отелло идут ровным, цельным и нерушимым шагом — crescendo и разрешаются дружными гармоническими аккордами* (И. А. Гончаров, 1875)*;*

(44) *Благодаря этим зонтам, пассажирам* [P] *есть возможность* [Y] *укрыться как от палящих лучей солнца, так и от дождя.* («Московский листок», 1902)*;*

(45) *Сейчас она в комнатушке при чьей-то конторе, сложена на полу и добраться до чего-либо мне* [P] *нет никакой возможности* [Y] (Р. Унгерн, 1926–1938)*;*

(46) *— Многое покажется, когда человеку* [P] *есть желание* [Y] *пить* (А. Платонов, 1929)*;*

(47) *Мне* [P] *нет выбора* [Y], *мой отец решительно не хочет, чтоб я шел в светское звание* (А. И. Герцен, 1857).

(48) *Я знал, что ты будешь колебаться, и потому неделю назад послал гонца к Юстиниану с письмом, в котором сообщил ему, что ты имеешь возможность без боя получить корону готов и Италии. Теперь тебе* [P] *нет выбора* [X] (Ф. Дан, Борьба за Рим, 1876, пер. с немецкого Д. И. Котляр, 1897)*;*

(49) *Ожиревшим камергерам* [P] *нет сил* [Y] *поспевать за ледащим и легким государем* (В. Шишков, 1934–1939)*;*

(50) *И мне* [P] *нет сил* [Y] *научить, вразумить себя — так грубы мои чувства, спеленан мой ум, в слухе звездные звуки — я не слышу себя, я не вижу себя!* (В. Ф. Одоевский, 1837)*;*

(51) *Охотнику* [P] *нет надежды* [Y] *выпутаться из ужасных когтей: он подает отчаянный сигнал, и лес с трепетом повторяет его* (И. И. Лажечников, 1838)*;*

(52) … *вся Россия в то время была, как тенетами, покрыта банками, так что ни одному зайцу* [P] *не было надежды* [Y] *проскочить, не попав головой в одну из петель* (М. Е. Салтыков-Щедрин, 1872).

Не вызывает сомнения, что во всех этих случаях современный редактор заменил бы дательный падеж на предложную группу типа *у*+S,род.[9] Появление дательного падежа в современном тексте (53) кажется неуместным анахронизмом или просто ошибкой:

(53) **И мне нет возможности поехать в Харьков* (Л. Спиридонова, «Наш современник», 2003).

Впрочем, примеров такого рода в современных текстах буквально единицы.

Эволюция дательного падежа актанта в рассматриваемых конструкциях весьма ярко проявляется в употреблении библейской фразы

(54) *Мне* [P] $\varnothing_{\text{БЫТЬ,наст}}$ [X] *отмщение* [Y]*, и аз воздам*.

Эта библейская фраза (Рим.12:19), хорошо известная и как эпиграф к «Анне Карениной» Л. Н. Толстого, вызывает многочисленные комментарии из-за неоднозначности ее современной интерпретации. Многим современным читателям кажется естественной интерпретация дательного падежа *мне* как адресата отмщения (ср., например, соответствующую дискуссию на сайте Грамота.ру (http://www.gramota.ru/spravka/hardwords/25_280). Эта фраза вошла в обиход из синодального перевода Нового Завета (где она выглядит как *Мне отмщение, Я воздам*). Этот перевод, вероятно, является калькой с греческого оригинала, где также присутствовала форма дат. п. εμοί ‘мне’[10], хотя, возможно, ее использование свидетельствует о том, что для переводчика субъектный дательный падеж был еще приемлем. Показательно, что в современных переводах Библии на другие языки текст этого выражения выглядит совершенно однозначно, ср. англ. перевод *Vengeance is mine. I will repay* или польск. *Pomsta do mnie należy, Ja odpłacę*. Надо сказать, что неоднозначность устранена и в более современных русских переводах Нового завета: *Отмщение — Мое, и Я воздам* (Современный русский перевод, НБО 2011).

---

[9] Курьезным подтверждением этого предположения является опубликованный на lib.ru вариант книги Ф. Дана, где вместо *Теперь тебе нет выбора,* как в (47), стоит *Теперь у тебя нет выбора* (http://www.lib.ru/INOSTRHIST/DAN/rim.txt_Piece100.06)

[10] В свою очередь, это место Нового Завета содержит отсылку к ветхозаветному *У Меня отмщение и воздаяние* (Втор.32:35, Синодальный перевод).

Разумеется, данная тенденция не является всеобщей. Для некоторых предикатных слов дательный падеж субъекта легко варьирует с конструкцией *у*+S,род (ср. выше примеры (39–41), для других он даже является предпочтительным (*мне нет покоя ни днем ни ночью* при сомнительности или редкости ?*у меня нет покоя ни днем ни ночью*). Для полной количественной оценки этой тенденции и наличия/отсутствия ее семантической мотивированности требуется дополнительное исследование, которое авторы намерены провести в будущем.

Литература

1. *Апресян* и др. 2010 — В. Ю. Апресян, Ю. Д. Апресян, Е. Э. Бабаева, О. Ю. Богуславская, И. В. Галактионова, М. Я. Гловинская, Б. Л. Иомдин, Т. В. Крылова, И. Б. Левонтина, А. В. Птенцова, А. В. Санников, Е. В. Урысон. Проспект активного словаря русского языка. Отв. ред. акад. Ю. Д. Апресян. М.: Языки славянских культур, 2010.
2. *Апресян* и др. 2014 — В. Ю. Апресян, Ю. Д. Апресян, Е. Э. Бабаева, О. Ю. Богуславская, И. В. Галактионова, М. Я. Гловинская, Б. Л. Иомдин, Т. В. Крылова, А. А. Лопухина, И. Б. Левонтина, А. В. Птенцова, А. В. Санников, Е. В. Урысон. Активный словарь современного русского языка. А–Г. Отв. ред. акад. Ю. Д. Апресян. М.: Языки славянских культур, 2014 (в печати).
3. *Арутюнова* 1976 — Н. Д. Арутюнова. Предложение и его смысл: Логико-семантические проблемы. М.: Наука, 1976.
4. *Богуславский* 2008 — И. М. Богуславский. Актантное поведение адвербиальных дериватов. // Динамические модели: Слово. Предложение. Текст. М.: Языки славянских культур. С. 110–129.
5. *Иомдин, Иомдин* 2013 — Л. Л. Иомдин, Б. Л. Иомдин. Отрицание и валентности в русском языке (по корпусным данным). // Труды международной конференции «Корпусная лингвистика — 2013). СПб.: С.-Петербургский гос. университет, С. 281–291.
6. *Boguslavsky* 2009 — I. Boguslavsky. Enlarging the Diversity of Valency Instantiation Patterns and Its Implications. Logic, Language, and Computation. Lecture Notes in Computer Science Volume 5422, 2009, pp. 206–220.
7. *Iomdin, Iomdin* 2011 — L. Iomdin, B. Iomdin (2011). Valency Ambiguity Interpretation: What Can and What Cannot be Done // Proceedings of the 5th International Conference on Meaning-Text Theory (MTT'2011). Barcelona, September 8–9, 2011. P. 108–119.
8. *Iomdin, Iomdin* 2014 — L. Iomdin, B. Iomdin (2014). Negation and Valencies of Russian Verbal Predicates // Meaning-Text Theory: Current developments. Wiener Slawistischer Almanach, Sonderband 85. (Ed. Valentina Apresjan, Boris Iomdin et al.) München: Kubon&Sagner, 2014 (in print).

References

1. *Apresjan Ju. D.* (ed.). (2010), Prospect of the Active Dictionary of Modern Russian. [Prospekt Aktivnogo slovarja sovremennogo russkogo jazyka]. Moscow.
2. *Apresjan Ju. D.* (ed.). (2014), Active Dictionary of Modern Russian. A–G [Aktivnyj slovar' sovremennogo russkogo jazyka. A-G]. Moscow (in print).
3. *Arutjunova N. D.* (1976), The sentence and its meaning: logical and semantical issues [Predlozhenie i ego smysl: Logiko-semanticheskie problemy]. Moscow.
4. *Boguslavsky I. M.* (2008). The actant behaviour of adverbial derivatives [Aktantnoe povedenie adverbial'nyx derivatov]. In: Dinamicheskie modeli: Slovo. Predlozhenie. Tekst. Moscow.
5. *Boguslavsky I. M.* (2009). Enlarging the Diversity of Valency Instantiation Patterns and Its Implications. In: Logic, Language, and Computation. Lecture Notes in Computer Science Volume 5422, 2009.
6. *Iomdin L. L., Iomdin B. L.* (2011). Valency Ambiguity Interpretation: What Can and What Cannot be Done. In: Proceedings of the 5th International Conference on Meaning-Text Theory (MTT'2011). Barcelona, September 8–9, 2011. P. 108–119. ISBN 978-84-615-1716-9.
7. *Iomdin L. L., Iomdin B. L.* (2013). Negation and valencies in Russian language (a corpus-based study) [Otricanie i valentnosti v russkom jazyke (po korpusnym dannym)]. In: Proceedings of the International conference on corpus linguistics "Korpusnaja lingvistika — 2013". ISBN 978-5-8465-1335-8. St.-Petersburg.
8. *Iomdin L. L., Iomdin B. L.* (2014). Negation and Valencies of Russian Verbal Predicates. In: Meaning-Text Theory: Current developments. Wiener Slawistischer Almanach, Sonderband 85. (Ed. Valentina Apresjan, Boris Iomdin et al.) München: Kubon&Sagner, 2014 (in print).

THE IMPACT OF MORPHOLOGY PROCESSING QUALITY ON AUTOMATED ANAPHORA RESOLUTION FOR RUSSIAN

Ionov M. (m.ionov@corp.mail.ru)
Mail.ru Group, Moscow, Russia

Kutuzov A. (andrey.kutuzov@corp.mail.ru)
Mail.ru Group, National Research University Higher School of Economics, Moscow, Russia

The paper deals with the problems of creating and tuning a system of automated anaphora resolution for Russian. Such a system is introduced, combining rule-based and machine learning approaches. It shows F-measure from 0.51 to 0.59. Freeling serves as an underlying morphological layer and an account of its quality is given, with its influence on anaphora resolution workflow. The anaphora resolution system itself is available to download and use, coming with online demo.

Keywords: anaphora resolution, morphology processing, machine learning, antecedent

1. Introduction

In this paper we describe **An@phora**—the system for automated pronominal anaphora resolution in Russian texts. The system was built as a participant of anaphora resolution systems evaluation forum to be held at the conference 'Dialog—2014'. It combines rule-based and machine learning approach to achieve better quality.

Anaphora (and coreference in general) resolution is crucial to many natural language processing applications, including dialog agents, machine translation, question answering systems, and a lot more. At the same time, Russian natural language processing community lacks open tools to accomplish this task. Unfortunately, published reports on automated Russian anaphora resolution (see the section 2) are few and most of the time not extensively documented with regard to precision and recall of approaches used. What's even more disappointing, tools themselves are not published. To our knowledge, up to now there is no open and available system for Russian anaphora resolution.

Thus, we addressed not only the task of constructing anaphora resolution machine itself, but also of making it publicly available under an open-source license. The system is implemented in Python, and is free to download and use[1]. Also, live demo is available[2], using Brat on-line markup system [Stenetorp et al 2012].

---

[1] https://github.com/max-ionov/russian-anaphora

[2] http://ling.go.mail.ru/anaphora

At the same time, it should be stressed that as in many other areas of natural language processing, quality of underlying stages of linguistic analysis is crucial for performance of the system. In the case of anaphora resolution, of great importance are tools which provide tokenization, sentence splitting, part-of-speech tagging and morphological analysis in general. Our system was based on open-source set of linguistic analysis tools Freeling [Padro et al 2012]. Though in general it showed satisfactory results, we had to fix a number of mistakes, described below. With increase in pre-processing performance, anaphora resolution performance grew accordingly.

The paper is organized as follows. In the section 2 we describe previous work in the field. In the section 3 anaphora resolution machine itself is presented. This section falls into two sub-sections, related to rule-based and machine-learning based modules of the machine. The next section evaluates the performance of the system. We describe typical errors both because of Freeling and because of incompleteness of our algorithm. Then we show how precision and recall measures change with fixing Freeling errors and with various experimental settings. Hybrid approach combining rules and machine learning is presented and its superior performance comparing to 'pure' approaches is shown. In the last section we conclude and propose future work.

2. Related work

Anaphora resolution for English is a well-developed field of natural language processing. First attempt was made in 1964 in an algebra problem-solving system STUDENT ([Bobrow 1964]). Since then the field saw much research.

Typically, anaphora resolution process consists of two steps: first, for each anaphor in the input text, a list of potential antecedent candidates is created. Second, the system decides which of the candidates is the most probable antecedent. Systems can be classified by the way they choose candidates. There are two dimensions of this distinction: types of features for choosing ("restrictional" or "preferential") and methods for choosing using sets of features (traditional rule-based or using machine learning algorithms). "Restrictional" methods are based on discounting candidates which do not satisfy features whereas "preferential" give more preference to those candidates that do satisfy them. An example of restrictional feature is number or gender agreement: candidates which do not agree with anaphoric expression are discarded. Syntax-oriented approaches, for example, Hobbs' algorithm ([Hobbs 1976]), use restrictional approach. An example of preferential feature is centering—giving preference to the most salient (focused) candidate. Detailed though a little outdated overview of anaphora resolution systems for English, features and approaches can be found in [Mitkov 1999]. In a recent evaluation of anaphora resolution for English best system performed with 73.94% F-measure ([Delmonte et al. 2006]).

Anaphora resolution for Russian is not so well-developed as for English. In [Tolpegin et al. 2006] a machine learning approach for third-person pronoun anaphora resolution is presented. Resolution was treated as a classification task, solved using Support Vector Machines. Three types of features were used for classification: distance, positional features and morphological features. The system performed with 62% precision. Unfortunately, recall of the system is unknown.

In [Malkovskiy et al. 2013] another pronominal anaphora resolution system is described. This system used syntactic features along with morphological and distance features and Random forest classification algorithm. Best result was achieved with all features—71% precision. Recall of the system is also unknown.

3. Anaphora resolution system

3.1. Rule-based approach

As stated above, anaphora resolution task falls into two stages: identification of the anaphoric pronoun and identification of its antecedent. First, one has to decide which lexemes are possible anaphoras.

In **An@phora** project we limited ourselves to the following pronouns, loosely separated into three groups: '**personal pronouns**' (*он, она, оно, они, его, её, их, мой*), '**reflexives**' (*себе, свой*), **'relatives'** (*который*). We dropped '*твой*' and '*тот*' from analysis because of their highly discursive nature: the choice of antecedent for these pronouns most of the time heavily depends on deep dialog structure. Moreover, in Russian their antecedent is often only inferred and not expressed by any particular word or multi-word expression. The training set provided by evaluation forum organizers lacks chains with anaphoric "твой" in any form, so it would be difficult to evaluate results even if we decided to handle this pronoun. It should be noted that this is not true for pronoun "мой", which often possesses proper antecedent, especially in the first person narratives.

Antecedent identification within the rule-based module is performed as follows. While reading the given text, we store all the words and noun phrases together with their morphological features. When this stack outgrows a given length (in words), it is shortened from the left to match the threshold. So, this 'analysis window' constantly moves along the text. It allows us to limit antecedent choice to only nearest candidates and not to confuse the system with candidates located far from the anaphoric expression. One can think about 'analysis window' as a kind of shallow salience detector. Experiments showed optimal length of analysis window for Russian texts to be around 23 words; see below.

Upon finding an anaphoric pronoun, the system looks to the left from it in the search of a noun phrase subject to specific constraints. Thus, our anaphora resolution system rule-based module can be classified as simple restrictional one. We presuppose that in most cases the nearest noun phrase abiding to these constraints is the antecedent.

Exact constraints are different for pronoun groups and for some separate pronouns. Simplified example for personal pronoun *она* ('she') will look like: search through all noun phrases within the analysis window. If singular feminine noun phrase in Nominative case[3] denoting animate object is found, consider it to be an antecedent and create an anaphoric chain. If no such noun phrase is found, take the

[3] Our experiments proved that Nominative noun phrases are preferred antecedents for personal pronouns, at least in the provided training set. Introducing this rule increased precision of the anaphora detector.

nearest singular feminine noun phrase and create a chain with it. If no suitable noun phrase is found, consider that the current pronoun is not linked to any antecedent.

Other pronouns have additional peculiarities and constraints. E.g., possessive pronouns of the first person search for their antecedent among first person pronouns, not among noun phrases, reflexives search among both of them, relatives check that there is a comma between anaphoric expression and antecedent, etc.

Despite its generally satisfactory performance (see evaluation section) and unmatched computation speed, rule-based system inherently suffers from its over-simplicity. It is difficult or even impossible to construct all combination of rules manually. Thus, a version of anaphora resolution machine using machine learning approach was designed.

3.2. Machine learning approach

To employ machine learning algorithms we considered anaphora resolution as a classification task: for each anaphoric expression we created a list of candidates and the classifier would label each of them as a possible antecedent or not. We used Random Forest as the main classification algorithm, mainly because it allows ranking features importance. We deliberately used excessive list of features to analyze which make the most contribution in classification. Moreover, Random Forest outperformed most of the popular algorithms (including SVM) for our task in synthetic tests.

As a preprocessing step we performed morphological analysis and simple noun phrase detection. Classifier was implemented using Scikit-learn library ([Pedregosa et al. 2011]). For each anaphoric expression the most probable candidate was returned if its probability was greater than threshold 0.3.

We used the following features for classification:

1. Length of the candidate group in characters
2. Length of the candidate group in words
3. Distance between pronoun and the candidate in characters
4. Distance between pronoun and the candidate in words
5. Distance between pronoun and the candidate in groups
6. Grammatical number of the candidate
7. Grammatical number of the pronoun
8. Do numbers of the candidate and the pronoun agree?
9. Grammatical case of the candidate
10. Grammatical case of the pronoun
11. Do cases of the candidate and the pronoun agree?
12. Is the candidate a proper name?
13. Number of the occurrences of candidate in the text
14. Pronoun type
15. Pronoun itself

Most of these features are fairly standard for this task. Features 12 and 13 are simple salience features. They and distance features are shown to be important (for example, in [Malkovskiy et al. 2013]). Some features, like case agreement, were added without any prior knowledge whether they are helpful or not, to determine their importance.

4. Experiments and evaluation

Evaluation of **An@phora** performance was made against the training set (gold standard), provided by evaluation forum organizers. The set consisted of 92 texts in Russian (69,282 words and 473,445 characters). It contained 2,141 annotated anaphoric chains. As a baseline, we constructed a simple algorithm which links anaphoric expressions to the nearest noun phrase to the left (in the case of reflexives, the nearest pronoun is chosen if it occurs first). This baseline algorithm performed with the following results:

Precision: **0.372**
Recall: **0.357**
F-Measure: **0.364**

4.1. Rule-based system performance

General performance of our system heavily depends on the choice of analysis window length. Our experiments showed that the length of 23 words is optimal with regard to F-measure. After reaching the limit of 23–25 words, analyzer performance quickly degrades, as seen on Fig. 1.

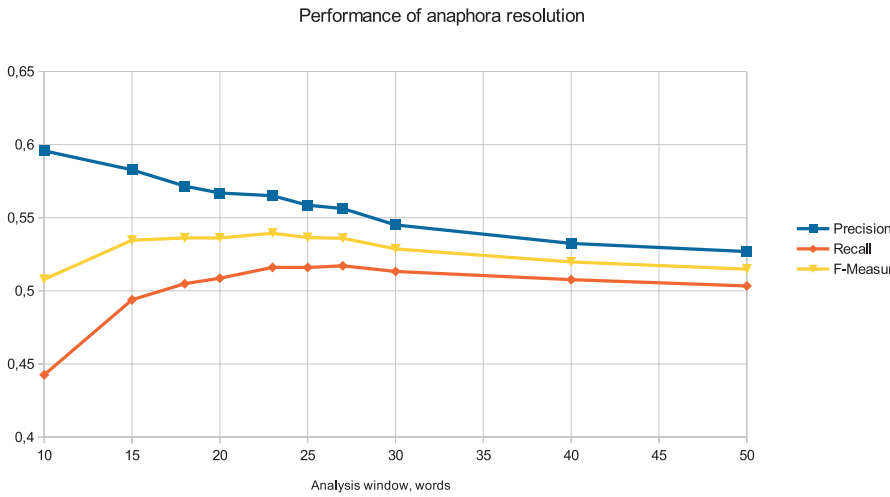


Fig. 1. Performance of anaphora resolution depending on analysis window length

At this optimal length the rule-based module reached **precision 0.565, recall 0.516 and F-measure 0.539** against gold standard, with F-measure 50% higher than the baseline. Attempts to use noun phrases as measure of analysis window length (instead of words) showed worse results with precision only 0.528 and recall 0.494, as well as measuring analysis window in characters (precision 0.47 and recall 0.45 at best). Thus, number of words is a finer setting.

Most frequent errors are related to:

1. Proper names, which are sometimes incorrectly analyzed by Freeling (wrong gender and number).
2. Incorrect choice between inanimate noun in Nominative case and animate noun in some other case. Cf. the expression *'У спутницы Олдмэна на лице траурная вуаль, она несет свои глаза на подносе.'* Our machine links *'она'* to *'траурная вуаль'*, as it is the nearest noun phrase, and additionally Nominative. *'Спутницы'* would win the contest only if it were Nominative, but it is not. Experiments with lending animate nouns bonuses independent of their case degraded performance.
3. Antecedent beyond the limits of analysis window. It is mostly found in encyclopedic articles with a lot of impersonal sentences, whose real subject is the article headword, and affects pronouns like *'себя'* and *'собой'*.
4. Cases of cataphora. Our system does not detect it, and finds incorrect antecedents to the left of the anaphoric expressions, when they are in fact on the right.
5. Direct speech (cases when our system links anaphoric expression to a noun group inside quotes). As in the above mentioned case with pronoun *'твой'*, proper handling of such cases demands processing of dialog structure, and we consider this to be future work.

4.2. Machine learning system performance

Evaluation of machine learning (ML) approach was performed on two random subcorpora as test sets while the rest of gold standard was used as a training set. This was done in order not to over-fit classifier using the same data for training and testing. Size of subcorpora are presented in the table 1.

Table 1. Size of subcorpora for ML system evaluation

	Test set, texts	Train set, texts
Subcorpus 1 (further S1)	5	106
Subcorpus 2 (further S2)	13	95

F-measure of machine learning (ML) resolution is generally worse than for rule-based (RB) system, however, precision is consistently higher. The results are presented in the table 2.

Table 2. Comparison of ML and RB systems performance on subcorpus S1 and subcorpus S2

	Precision	Recall	F-measure
Rule-based, S1	45.02%	46.58%	45.79%
Machine Learning, S1	52.54%	43.51%	47.60%
Rule-based, S2	63.57%	56.45%	59.80%
Machine Learning, S2	65.11%	45.67%	53.69%

Analysis of feature importances in both models shows this order for features with importance > 0.05:

1. Distance in characters
2. Distance in words
3. Distance in groups
4. Length of candidate in characters
5. Pronoun
6. Number of the occurrences of candidate in text
7. Case of the candidate
8. Type of the pronoun

As we can see, distance appears to be the most important feature, whereas number and case agreement are much less important. Interestingly, distance in characters appears to be more important than distance in words (0.193 and 0.121, respectively). This needs further analysis because this result is far from obvious.

4.3. Influence of morphologic processing on system performance

It turns out that performance of preliminary NLP steps, such as morphological analysis, has crucial influence on performance of anaphora resolution. E.g., we discovered several inconsistencies in Freeling handling of Russian pronouns (supposedly, not Freeling itself should be blamed for that, but the corpus on which its Russian module had been trained). Fortunately, Freeling is very flexible and allows to fine-tune its morphologic analysis model. Among others, we had to fix number and case probabilities for *'его'* and add missing gender annotation to almost all personal pronouns: more than fifteen corrections total.

We compared performance of our rule-based anaphora resolution machine with original and fixed Freeling. The results are given in the table 3.

Table 3. Performance increases after fixes to morphology processing of anaphoric pronouns

	Original Freeling	**Fixed Freeling**
Precision	0.493	0.565
Recall	0.424	0.516
F-Measure	0.456	0.539

This dramatic difference comes as no surprise. We extensively use number, gender and case features of anaphoric expressions to check their agreement with antecedents. Thus, insufficient or outright incorrect morphological processing directly influences anaphora resolution performance.

It should be noted that of course Freeling errors are not limited to pronouns. Occasionally it wrongly detects noun case or gets stuck on nominalized adjectives. We experimented with a text from gold standard 1,389 words long. Our machine

detected anaphoric chains in this text with precision 0.3, recall 0.28 and f-measure 0.29. However, after we manually fixed Freeling output for all words of the text, precision raised to 0.31, recall to 0.3 and f-measure to 0.3. Among others, manual post-processing fixed treating proper surname *'Одиноков'* as genitive plural and let pronoun *'их'* link to correct antecedent instead of this.

At the same time, all in all we had to make only 86 corrections to the annotation of the text containing 1,389 words. Thus, only 6% of Freeling output demanded any manual intervention (and a lot of them only slight one, like adding animation property). We consider it a sufficient degree of quality. However, we also plan to return all our corrections to Freeling developers to be incorporated in the next release.

4.4. Hybrid approach

Machine learning approach shows higher precision than rule-based approach but lower recall, thus we created a hybrid system to improve both results using advantages of each method. Low recall with high precision means that if the system returns a result it is confident about it. So we integrated a new stage in our rule-based pipeline: for each pronoun we tried to predict antecedent with ML. If it couldn't predict, we used rule-based approach. This method showed improvement in average: when rule-based approach shows low F-measure, hybrid one improves the result drastically, when the rule-based approach shows high F-measure, hybrid one may lower overall results but not critically. See table 4 for comparison.

Table 4. Comparison of ML, RB and hybrid systems' performance on subcorpora S1 and S2

	Precision	**Recall**	**F-measure**
Rule-based, S1	45.02%	46.58%	45.79%
Machine Learning, S1	**52.54%**	43.51%	47.60%
Hybrid, S1	49.49%	**53.72%**	**51.52%**
Rule-based, S2	63.57%	**56.45%**	**59.80%**
Machine Learning, S2	**65.11%**	45.67%	53.69%
Hybrid, S2	62.22%	**56.46%**	**59.20%**

Thus, hybrid approach seriously outperforms both rule-based and machine learning ones on S1 subcorpus and is almost on a par with rule-based algorithm on S2 subcorpus.

5. Conclusion and future work

We presented **An@phor**a—a system for automated anaphora resolution in Russian texts. It is freely available under open-source GPL license and can be tested through online demo.

Our approach includes using Freeling as an underlying morphologic analysis layer and a combination of rules and machine learning model to ensure better anaphora resolution quality. Our separate rule-based module, tested against training set provided by evaluation forum organizers, showed **F-measure of 0.539 with precision 0.565 and recall 0.516.** General hybrid module, tested on two different subcorpora from the training set, showed **F-measure from 0.51 to 0.59**. General improvement of F-measure in comparison to a simple baseline algorithm is 40% to 62%.

At the same time, there is still room for future improvement. We plan to train our classifier on a larger corpus, as 70 thousand words from gold standard is clearly not enough. Rule-based module can also be improved to consider direct speech issues, cataphora and sentences interaction. Finally, handling of "твой" and "тот" anaphoric expressions should be implemented.

6. Acknowledgments

The study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) in 2014.

References

1. *Delmonte, R., Bristot, A., Piccolino Boniforti, M. A., and Tonelli, S.* (2006). Another Evaluation of Anaphora Resolution Algorithms and a Comparison with GETARUNS' Knowledge Rich Approach, In Proc. of ROMAND 2006, Trento, pp. 3–10.
2. *Hobbs, Jerry R.* (1976) Pronoun Resolution. Research Report 76-1, Department of Computer Sciences, City College, City University of New York. August 1976.
3. *Lluís Padró and Evgeny Stanilovsky* (2012), FreeLing 3.0: Towards Wider Multilinguality, Proceedings of the Language Resources and Evaluation Conference (LREC 2012), Istanbul, Turkey.
4. *Mitkov R.* (1999) Anaphora resolution: the state of the art.—School of Languages and European Studies, University of Wolverhampton
5. *Pedregosa* et al. (2011) Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825–2830, 2011.
6. *Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou and Jun'ichi Tsujii* (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. Proceedings of the Demonstrations Session at EACL 2012.
7. *Tolpegin P. V., Vetrov D. P., Kropotov D. A.* (2006) Algorithm for machine-learning based automated resolution of third person pronominal anaphora [Алгоритм автоматизированного разрешения анафоры местоимений третьего лица на основе методов машинного обучения]. Proceedings of Dialog-2006 conference (Moscow, Bekasovo, 2006)
8. *Malkovsky M. G., Starostin A. S., Shilov I. A.* (2013) A method for pronominal anaphora resolution in the course of syntactic analysis [Метод разрешения местоименной анафоры в процессе синтаксического анализа] Proceedings of Sworld conference, pp. 41–49.

МЕТОДЫ РАЗРЕШЕНИЯ МЕСТОИМЕННОЙ АНАФОРЫ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

Каменская М. А. (ma_kamenskaya@mail.ru)
Российский университет дружбы народов, Москва, Россия

Храмоин И. В. (hramoin@isa.ru),
Смирнов И. В. (ivs@isa.ru)
Институт системного анализа Российской академии наук, Москва, Россия

Ключевые слова: разрешение анафоры, машинное обучение, метод опорных векторов, деревья решений, семантические роли

DATA-DRIVEN METHODS FOR ANAPHORA RESOLUTION OF RUSSIAN TEXTS

Kamenskaya M. A. (ma_kamenskaya@mail.ru)
Peoples' Friendship University of Russia, Moscow, Russia

Khramoin I. V. (hramoin@isa.ru),
Smirnov I. V. (ivs@isa.ru)
Institute for Systems Analysis of RAS, Moscow, Russia

The paper considers two data-driven methods for anaphora resolution of Russian texts. These methods are based on machine learning with annotated corpora and using no additional information except linguistic features. The first method uses Support Vector Machine as learning and classifying algorithms, the second method uses Decision Tree inducer. We evaluate the performance of the methods with several feature sets and corpora. Feature sets included morphological, syntactic and semantic features. In this paper we also evaluate how semantic features, namely semantic roles, impact the performance of anaphora resolution in Russian. We used our manually annotated corpus as well as a corpus provided by the organizing committee of the forum for the evaluation of linguistic text analysis systems, an event of Dialogue 2014. Experiments showed that precision of SVM is higher on experimental data for almost all cases. It was shown that semantic features enhance the performance of the methods for anaphora resolution of Russian texts. We have also calculated the optimal distance between the anaphor and the hypothetic antecedent and used it in our methods.

Key words: anaphora resolution, machine learning, support vector machine, decision trees, semantic roles

1. Introduction

Anaphora resolution is one of the core problems of natural language processing. Methods for anaphora and coreference resolution are used in systems for machine translation, information retrieval, information extraction and others. The problem of anaphora resolution is widely researched for English and other European languages. For Russian this problem had been solving by different researchers but until Dialog-2014, there had been no objective evaluation of methods for anaphora resolution of Russian.

In this paper we solve two tasks: the first one is to evaluate two simple data-driven methods for anaphora resolution of Russian which use no additional information except linguistic features. These methods based on machine learning with several feature sets using annotated corpora. The second task is to investigate how semantic features, namely semantic roles, influence performance of anaphora resolution of Russian.

We deal only with pronominal anaphora resolution and compare two approaches—statistical one, based on Support Vector Machine, and inductive method for Decision Tree construction. As learning and testing data sets, we used two annotated corpora: the first is our own, the second one was provided by organizers of Dialog-2014 parsers evaluation task. Feature set included morphological, syntactic and semantic features, obtained from semantic parser developed in ISA RAS [Osipov et al., 2008].

In section 2 related works are reviewed, section 3 describes feature sets, section 4 describes methods and section 5 presents experiment results. Section 6 presents conclusion and future work.

2. Related works

Research in automatic pronominal anaphora resolution for English started in the 70th. The first methods and systems by Winograd, Wilks, Hobbs [Mitkov, 1999] operated with the rules relying mostly on syntactical information; in addition, encyclopedic knowledge was also widely applied. In the 80th the tendency of combining different features, which had been used separately before, appeared. The papers of E. Rich and S. LuperFoy, J. Carbonell, R. Mitkov described the algorithms that combined agreement of gender and number, syntactic and semantic relations. In the 90th rule-based algorithms were replaced with statistical data-driven algorithms. I. Dagan and A. Itai, Connolly, Burger used the machine learning methods for anaphora resolution for the first time. For learning more about works of that time, one can turn to the paper of R. Mitkov [Mitkov, 1999]. The author discusses the history of the problem, traditional methods for anaphora resolution and characterizes the known computer systems.

Modern approaches are based on automatic learning using annotated corpora. They combine traditional linguistic methods with statistical methods and use different types of knowledge: morphologic, syntactic, semantic, and additional information, such as thesauri. A lot of interesting ideas and methods were represented at the CoNLL-2011 Shared Task [Pradhan et al., 2011]. The system [Lee et al., 2011] based on combination of multi-pass sieves, which incorporate lexical, syntactic, semantic, and discourse information, showed the best results. A variety of data sets available

for learning coreference resolution systems for English (see, for example [MUC-6 data set, 1995]) provides progress of research in this field.

Anaphora resolution for Russian is less experimentally researched. In [Kibrik, 1996] author discusses theoretical aspects of the anaphora phenomenon for Russian language and describes the series of linguistic features, reflecting nature of anaphora. One of the latest Kibrik's papers [Kibrik et al., 2013] is rather informative. The works of Tolpegin [Tolpegin, 2006], [Tolpegin et al., 2006] are also well known. The author proposes algorithm for construction of statistic model for pronominal anaphora resolution in Russian texts using machine learning methods. In paper [Abramova et al., 2011] authors describe in detail principles of anaphoric relations detection in different sentences and situations, which they use for the analysis of rules of socio-political texts coherence. The research of Mal'kovskij [Mal'kovskij et al., 2013] is one from the latest known papers for Russian. The work deals with the rule-based method for pronominal anaphora resolution, which uses analysis of words collocation. The core problem for Russian is absence of open data sets for learning coreference resolution methods and their evaluation.

There are the series of works that evaluate the influence of semantic knowledge on anaphora resolution quality. The papers [Ponzetto and Strube, 2006], [Kong et al., 2008], [Huang et al., 2009], [Zhou et al., 2001] demonstrate that using semantic roles as additional features improves anaphora resolution performance for English. The authors use data-driven methods but the approaches to selecting the groups of features and learning algorithms are different.

3. Feature set

We consider anaphora resolution problem as classification problem and solve it using machine-learning methods. The following features were used for leaning and classification:

Morphological and syntactic features:

1) gender, number, case, and animate of anaphor;
2) gender, number, case, and animate of antecedent;
3) comparison of anaphora's animate and antecedent's animate;
4) number of sentences between anaphor and antecedent;
5) number of words between anaphor and antecedent;
6) number of hypothetical antecedents between anaphor and antecedent;
7) number of nouns between anaphor and antecedent;
8) name of syntactic relation between anaphor and antecedent;

Semantic features:

9) semantic roles of anaphor;
10) semantic roles of antecedent;
11) combination of categorical semantic class of the head word of syntactic phrase, which contains anaphor as related word, and categorical semantic class of the head word of syntactic phrase, which contains antecedent as related word;

12) combination of categorical semantic class of the head word of syntactic phrase, which contains anaphor as related word, and categorical semantic class of the antecedent.

We use features 1–3, because we suppose that gender, number and animate of anaphor should agree with gender, number and animate of antecedent. Features 4–7 give information about distance between anaphor and antecedent in different scales. Feature 8 was proposed, because we guess that antecedent can be a part of fixed number of syntactic relations as a related word. We also expect antecedent to be a related component in verbal phrase. A word can be labeled with several semantic roles, because it can be an argument for different situations described in one sentence, especially in complex sentences. We use features 11–12, because we suppose that the categorical semantic class of noun can be combined with the fixed number of categorical semantic classes of verbs, as well as categorical semantic classes of verbs, which are associated with the same noun, are combined according to the special rules.

Features' values were obtained as a result of morphological, syntactic and semantic analysis of texts [Osipov et al., 2008]. Methods for semantic role labeling of Russian are described in [Smirnov et al., 2014]. Detailed lists of categorical semantic classes and semantic roles are presented in [Osipov, 2001]. We experimented with two feature sets: feature set FS-1 included features 1–8, feature set FS-2 included features 1–8 and semantic features 9–12.

4. Methods

Anaphora resolution is a task of detecting correct pairs "antecedent-anaphor". In our research, we deal only with personal, reflexive and demonstrative pronouns. The training set contains examples of correct and incorrect "antecedent-anaphor" pairs. Correct "antecedent-anaphor" pair contains correct hypothetic antecedent, incorrect "antecedent-anaphor" pair contains incorrect hypothetic antecedent. Hypothetic antecedent is a noun or pronoun for which anaphora has been already resolved. Hypothetic antecedent must be agreed with anaphor by number and gender. The distance in words between the anaphor and the antecedent should be not more than preliminary defined value that depends on corpus. Training example is presented as set of values of named features described in the previous section.

4.1. The algorithm of constructing training data set using annotated corpus

1. Find first annotated "antecedent-anaphor" pair in corpus.
2. Look for all nouns or pronouns for which anaphora has been already resolved, between the anaphor and the antecedent. Their number and gender must be agreed with anaphor's number and gender. The search area is limited by a predefined number of words.

3. All nouns and pronouns, which were found in the step 2 are incorrect hypothetic antecedents.
4. If correct antecedent is not in a search area, it will not be added to the training set.
5. Do steps 1–4 for each annotated example.

For training and classifying correct/incorrect pairs we used Support vector machine method (SVM) [Chang and Lin, 2014] and decision tree method [University of Waikato, 2014] with REPTree learner.

4.2. The algorithm for anaphora resolution

1. Find first anaphor, for which antecedent has not already been found. If anaphor has not been found, algorithm finishes.
2. Look for all nouns or pronouns for which anaphora has been already resolved, between the anaphor and the antecedent. Their number and gender must be agreed with anaphor's number and gender. The search area is limited by a predefined number of words.
3. Add them to hypothetic antecedents' set.
4. Assign to each pronoun in hypothetic antecedents' set categorical semantic class of its' antecedent.
5. Calculate the probability of each hypothetic antecedent to be correct antecedent using classification method.
6. Choose antecedent which has the highest probability and link it with the concerned anaphor. Go to step 1.

Area for searching hypothetic antecedent is limited by the number of words in step 2, because anaphor usually refers to nearest hypothetic antecedent. This value has been calculated in our experiments.

5. Results of experiments

Experiments have been run on several manually annotated corpora. The first corpus CORPUS-1 contains 17 texts of the Moshkov's library and 34 texts of SynTagRus [Apresjan et al., 2005]. CORPUS-1 contains 910 "antecedent-anaphor" pairs. CORPUS-2 is the annotated corpus, provided as a training set by the organizing committee of the Dialogue-2014 forum for the evaluation of linguistic text analysis systems. CORPUS-2 contains 92 texts and 967 "antecedent-anaphor" pairs. CORPUS-3 is union of CORPUS-1 and CORPUS-2.

The results of the preliminary experiments showed that distance in words, which limit hypothetic antecedent's searching area, is one of the most important features. This feature has a significantly positive effect on precision and recall of automated anaphora resolution because rise in distance between hypothetic antecedent

and anaphor lowers probability, that the anaphor refers to this antecedent [Tolpegin et al., 2006]. Moreover, rise in that distance causes rise in number of hypothetic antecedents, which makes anaphora resolution more complicated, costly and long. This is the reason to find the optimal distance that limits hypothetic antecedent's searching area. Such distance should include correct antecedent most time and limit the number of hypothetic antecedents as much as possible.

The distributions of number of correct "antecedent-anaphor" pairs according to distance between antecedent and anaphor for each corpus are presented on figures 1–3. We calculated the optimal distance that covers 90% of correct "antecedent-anaphor" pairs for every corpus, using these distributions. This optimal distance is equal to 25 words for CORPUS-1, 14 words for CORPUS-2 and 18 words for CORPUS-3. Hypothetic antecedents were searched not farther than calculated optimal distance in both training and classifying process.

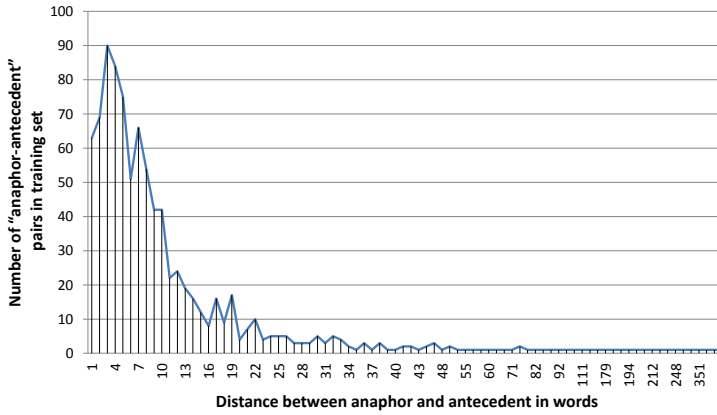


Fig. 1. Number of "antecedent-anaphor" pairs in relation on distance between anaphor and antecedent in CORPUS-1

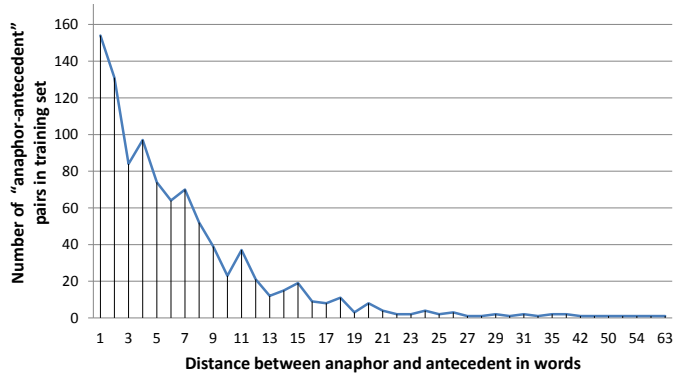


Fig. 2. Number of "antecedent-anaphor" pairs in relation on distance between anaphor and antecedent in CORPUS-2

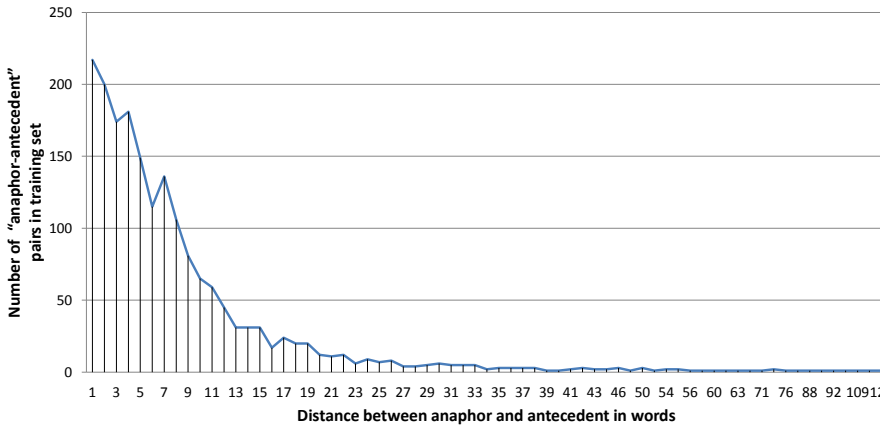


Fig. 3. Number of "antecedent-anaphor" pairs in relation on distance between anaphor and antecedent in CORPUS-3

We used the following metrics: SCORE-1 is a precision of recognition of both correct and incorrect "antecedent-anaphor" pairs (precision of classification of examples into two classes—correct or incorrect), SCORE-2 is the precision of finding correct antecedent for each anaphor. We use only CORPUS-2 as testing corpus to calculate SCORE-2. SCORE-2 represents actual precision of anaphora resolution and is most close to the task. We use ten-fold cross validation to calculate SCORE-1. Scores of SVM method with feature set FS-1 were chosen as a baseline.

The result of the first experiment on CORUS-1 is presented in table 1.

Table 1. Precision of anaphora resolution for different methods and feature sets on CORPUS-1 as training corpus

Feature set	SVM	REPTree
SCORE-1		
FS-1	0.811	0.773
FS-2	0.821	0.789
SCORE-2		
FS-1	0.473	0.484
FS-2	0.539	0.529

The result of the second experiment on CORUS-2 is presented in table 2.

Table 2. Precision of anaphora resolution for different methods and feature sets on CORPUS-2 as training corpus

Feature set	SVM	REPTree
SCORE-1		
FS-1	0.746	0.746
FS-2	0.771	0.747
SCORE-2		
FS-1	0.603	0.592
FS-2	0.61	0.609

The result of the third experiment on CORUS-3 is presented in table 3.

Table 3. Precision of anaphora resolution for different methods and feature sets on CORPUS-3 as training corpus

Feature set	SVM	REPTree
SCORE-1		
FS-1	0.766	0.634
FS-2	0.781	0.689
SCORE-2		
FS-1	0.571	0.548
FS-2	0.579	0.553

We have done 8 runs on the test corpus provided by the organizing committee of the Dialogue-2014 forum for the evaluation of linguistic text analysis systems. The methods were learned on CORPUS-3 with feature sets FS-1 and FS-2.

6. Conclusion and future work

The results of experiments showed the reasonable results for both simple methods. SVM exceeded DT by 0,1%–13,2% of precision for almost all runs on experimental data. Anaphora resolution using semantic features showed precision gain by 0,1%–6,6% for all methods and runs. Such values for precision gain in this case can be explained by the fact that semantic roles were assigned to both anaphor and antecedent in only 8% of anaphoric pairs in manually annotated learning corpus. The F-measure of semantic parser used for testing anaphora resolution is 75% with 67% of recall and 86% of precision, so precision gain for anaphora resolution is adequate and rather good. The best result on test corpus (61% of precision) was shown by SVM on CORPUS-2 with feature set FS-2.

Thus, experiments showed that semantic features enhance performance of methods for pronominal anaphora resolution of Russian texts. As a future work, we will extend feature set with extra-lingual information using several thesauri and enhance method for identifying hypothetic antecedents.

References

1. *Abramova N. N., Abramov V. E., Nekrasova E. V., Ross G. N.* (2011), Statistic analysis of social and political texts coherence [Statisticheskij analiz svjaznosti tekstov po obshchestvenno-politicheskoj tematike], Proceedings of the 13th All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" [Trudy 13j Vserossijskoj nauchnoj konferentsii "Èlectronnye biblioteki: perspektivnye metody i tehnologii, èlektronnye kollektsii"], Voronezh, pp. 127–133.
2. *Apresjan J. D., Boguslavskij I. M., Iomdin B. L., Iomdin L. L., Sannikov A. V., Sannikov V. G., Sizov L. L.* (2005), Syntactically and semantically annotated corpus of Russian language: Present state and perspectives [Sintaksicheski i semanticheski annotirovannyj korpus russkogo jazyka: sovremennoe sostojanie i perspektivy], National Corpus of Russian Language: 2003–2005 [Natsional'nyj korpus russkogo jazyka: 2003–2005], pp. 193–214.
3. *Chang C.-C., Lin C.-J.* (2014), LIBSVM—A Library for Support Vector Machines, available at: www.csie.ntu.edu.tw/~cjlin/libsvm/
4. *Huang Z., Zeng G., Xu W., Celikyilmaz A.* (2009), Accurate semantic class classifier for coreference resolution, Conference on Empirical Methods in Natural Language Processing, Singapore, pp. 1232–1240.
5. *Kibrik A. A.* (1996), Anaphora in Russian narrative discourse: A cognitive calculative account In B, Fox (ed.) Studies in anaphora, Amsterdam, pp. 255–304.
6. *Kibrik A. A., Dobrov G. B., Khudyakova M. V., Loukachevitch N. V., Pechenyj A.* (2013), A corpus-based study of referential choice: Multiplicity of factors and machine learning techniques, Text processing and cognitive technologies. Cognitive modeling in linguistics: Proceedings of the 13th International Conference, Corfu, pp. 118–126.
7. *Kong F., Li Y., Zhou G., Zhu Q., Qian P.* (2008), Using Semantic Roles for Coreference Resolution, International Conference on Advanced Language Processing and Web Information Technology, pp. 150–155.
8. *Lee H., Peirsman Y., Chang A., Chambers N., Surdeanu M., Jurafsky D.* (2011), Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task, *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task* (CONLL Shared Task '11), Stroudsburg, pp. 28–34.
9. *Mal'kovskij M. G., Starostin A. S., Shilov I. A.* (2013), Method of pronominal anaphora resolution in parallel with syntactic analysis [Metod razreshenija mestoimennoi anafory v protsesse sintaksicheskogo analiza], Perspective innovations in science, education, production and transport'2013 [Perspektivnye innovatsii v nauke, obrazovanii, proizvodstve I transporte'2013], available at: www.sworld.com.ua/index.php/ru/technical-sciences-413/informatics-computer-science-and-automation-413/20828-413-0615.

10. *Mitkov R.* (1999) Anaphora resolution: the state of the art, Working paper (based on the COLING'98/ACL'98 tutorial on anaphora resolution), available at: clg.wlv.ac.uk/papers/mitkov-99a.pdf
11. MUC-6 data set, (1995), available at: http://cs.nyu.edu/faculty/grishman/muc6.html
12. *Osipov G. S., Smirnov I. V., Tikhomirov I.* (2008), Relational–situational method for search and analysis of texts and its applications [Reljatsionno-situatsionnyj metod poiska i analiza tekstov i ego prilozhenija], Artificial Intelligence and Decision Making [Iskusstvennyj intellekt i prinjatie reshenij], (1), pp. 3–10.
13. *Osipov G. S.* (2011), Methods of artificial intelligence [Metody iskusstvennogo intellekta], FIZMATLIT, Moscow.
14. *Ponzetto S. P., Strube M.* (2006), Semantic role labeling for coreference resolution, Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations (EACL'06), Trento, pp. 143–146.
15. *Pradhan S., Ramshaw L., Marcus M., Palmer M., Weischedel R., Xue N.* (2011), CoNLL-2011 shared task: modeling unrestricted coreference in OntoNotes, *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task* (CONLL Shared Task'11), Stroudsburg, pp. 1–27.
16. *Smirnov I. V., Shelmanov A. O., Kuznetsova E. S., Hramoin I. V.* (2014), Semantic-syntactic analysis of natural languages. Part II. Method for semantic-syntactic analysis of texts [Semantiko-sintaksicheskij analiz estestvennyh jazykov Chast' II. Metod semantiko-sintaksicheskogo analiza tekstov], Artificial Intelligence and Decision Making [Iskusstvennyj intellekt i prinjatie reshenij], (1), pp. 95–108.
17. *Tolpegin P. V.* (2006), The new methods and algorithms of automated third person pronominal reference resolution of Russian texts, [Novye metody i algoritmy avtomaticheskogo razreshenija referentsii mestoimenij tret'ego litsa russkojazychnyh tekstov], Komkniga, Moscow.
18. *Tolpegin P. V., Vetrov D. P., Kropotov D. A.* (2006), Automated third person anaphora resolution algorithm on the basis of machine learning methods [Algoritm avtomatizirovannogo razreshenija anafory mestoimenij tret'ego litsa na osnove metodov mashinnogo obuchenija], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2006", [Komp'juternaja Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2006"], Bekasovo, pp. 504–507.
19. University of Waikato, (2014), Weka 3: Data Mining Software in Java, available at: www.cs.waikato.ac.nz/ml/weka/
20. *Zhou H., Li Y., Huang D., Zhang Y., Wu C., Yang Y.* (2011), Combining syntactic and semantic features by SVM for unrestricted coreference resolution, *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task* (CONLL Shared Task'11), Stroudsburg, pp. 66–70.

ПРАГМАТИЧЕСКИЕ АСПЕКТЫ ИНТЕРНЕТ-КОММУНИКАЦИИ: К РАЗРАБОТКЕ ЖАНРОВЫХ МОДЕЛЕЙ ВЕБ-САЙТОВ

Кононенко И. С. (irina_k@cn.ru)
Институт систем информатики им. А. П. Ершова
СО РАН, Новосибирск, Россия

Ключевые слова: интернет-жанр, веб-сайт, праксиологический параметр, коммуникативный параметр, жанровая структура, жанровый маркер

PRAGMATIC ASPECTS OF INTERNET COMMUNICATION: TOWARDS WEBSITES GENRE MODELS[1]

Kononenko I. S. (irina_k@cn.ru)

A.P. Ershov Institute of Informatics Systems, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

A two-level multifaceted genre classification is proposed to cover pragmatic aspects of communication on the Web. Genre categories of websites and genre types of site constituents (pages and structural blocks) are represented as vectors of relevant pragmatic features. Praxeological parameters (activity subject, beneficiary, product, environment) are involved to represent human activity that underlies communication and manifests itself in the site structure, content and form of site constituents. Communicative parameters encompass the hierarchy of communicative tasks (including anticipated reactions of the target audience), functionality of site constituents, and the affordances of communication channel (interactivity, multimodality, and dynamics of content). Functions of site constituents together with medium features are exemplified to determine genre types of pages. The type of a textual page corresponds to a certain genre schematic structure composed of content blocks. The extraction of genre schemata is possible using the so called genre markers (cue words and constructions) that are formalized as lexico-grammatical patterns provided with format conditions.

Key words: web genre, website, praxeological parameter, communicative parameter, genre schematic structure, genre marker

[1] The research was supported by Russian Foundation for Basic Research, project No 13-07-00422.

1. Introduction

Recently, special attention is paid to communication practices realized in the digital environment via numerous web pages and sites. These phenomena are investigated within new research fields such as Internet linguistics and digital genre studies [Santini et al. 2010, Shchipicina 2010]. At the same time the widespread communication on the Web stimulates site constructors to greater effort to support the web-based intercourse with technological and software solutions aimed at automatization of website development [Site Technologies Inc.]. Within this context the more general task is discussed: to support automatic generation of the website structure and design as well as certain elements of the site content.

Possible decision may be based on common features, relatively stable stylistic and compositional types observed in the wide variety of websites. All these features correspond to the classic definition of text genre, according to which genre is a type model for the speech unit construction [Bahtin 1986]. So, any website should be in compliance with some genre model that presents its "standard repeated genre form".

The wide range of modern approaches to web genres is represented in [Mehler et al. 2010]. A very useful summary is provided in introduction [Santini et al. 2010], whose authors emphasize limitations of purely topical classification of web resources. Really, presentations of the same topic may be essentially distinct, for example, the implantation problem described on the website of a dental office, as opposed to that discussed in the dental health forum. Genre categories should be topically neutral as much as possible, though specific genre forms may be more or less closely related to topics. Yandex catalog of Russian websites allows for the opposition of topic and genre by using the brunched and extensive topical hierarchy, along with the simple classification by information types (goods and services proposals, advice and instruction, reference works, forums, and events) [Yandex catalog]. A. A. Kibrik considers genres in close relation with functional styles, which reflect spheres of human activity [Kibrik 2009]. [Crowston et al. 2010] point out that genre is a medium for participation in a communicative act; so, in identifying and labeling genres "the gestalt of the various components of the communicative act" is to be captured.

An intrinsic multifaceted nature of web genre implies functional, formal (compositional and lexico-grammatical), and content aspects, which correspond to at least two levels of physical representation: website as a whole and site constituents (site section as a number of web pages, individual web page, and structural block, or move). Website is a manifestation of a communicative act addressed by the author to the target audience and performed as a part of some human activity. Site genre is a conventional way to perform the communicative act on the Web, so an attempt to systematize and formalize site generation requires considering two types of pragmatic parameters that may form the base of multifaceted classification of web genres:

- praxeological parameters, which are facets to classify spheres of human activities,—to represent activities context of communication (section 2);
- communicative parameters that represent communicative context proper: communicative tasks and characteristics of communicative situation brought about by the medium (channel) of communication (section 3).

- Thus obtained genre categories correlate with site genre models considered as patterns for structuring information with respect to several aspects, in particular, by putting it into compositional scheme and lexico-grammatical shape. In section 4 it is described how site genre pragmatics is reflected in the compositional scheme of a web page text. This structure is identified with the help of genre markers represented as lexico-grammatical patterns.

2. Praxeological Parameters

It is common to describe human activity as a process of purposeful interaction of a subject with an object, this process involving such components as a product (realized goal), resources, and conditions.

In the activities ontology all the concepts are considered as entities that are classified as Objects or Processes, on the one hand, and Agents and Non-Agents, on the other. The Process determines a number of Roles (role relations), each of which represents certain aspect of the process and puts the constraints on the potential role-fillers (participants). Agent entities, as opposed to Non-Agents, are conscious, volitional, and able to intentional activity, hence, they may fill the role of the activity Subject. Other roles that may be filled by Agentive entities are Object, Counter-agent, and Beneficiary. Non-Agent entities are physical, social, and mental objects and processes, as well as temporal and locative objects. Their roles in the activity process are Object, Product, Beneficiary, Time, Location, etc. Ontological concepts can be refined within some specified taxonomy. Concepts, both entities and relations, are characterized by attributes that describe conceptual properties.

Proceeding from these ontological considerations, a set of praxeological parameters has been introduced to build a multifaceted categorization of activities. Some of these parameters and corresponding genre based manifestations are exemplified below.

Subject: {*Individual, Group, Institution, State*}. *Group* can be refined as a *Social group* (particularly *Family*) or a *Community* (such as interest group). Different values of the parameter are reflected in the genre form: first person singular pronouns are specific to personal pages (personal communication) and quite impossible on the website of organization (institutional communication); first person plural pronouns are frequent on institutional commercial and advertising pages but not on the official site of the state authority.

Beneficiary: {*Individual, Group, Institution, Society*}. Benefactive relation corresponds to the participant supposed to make use of the activity results (Product). Different values distinguish benefits (and corresponding activities) that are used individually, collectively, or by the whole society.

Product Separability {*Goods, Services*}. It differentiates between *production of goods* and *rendering services*, as the process of rendering a service overlaps or coincides with consumption of its results. Inseparability can explain, for example, the introduction of constituent block "making an appointment" in the structure of services website.

Product Substantiality: {*Substantial, Unsubstantial*}. *Unsubstantial* benefits are mental or physical properties of a person (health, knowledge, etc.).

Product Type {*Material, Financial, Informational, Spiritual, Vital*}. The combination of "substantiality" and "type" values represent benefit varieties:

<*Substantial, Material*>—the results of *material production* (equipment, buildings, furniture) or *material services* (public utilities, freight services);

<*Unsubstantial, Vital*>—the state of physical or mental state as a result of *vital* (medical or recreational) *services*;

<*Substantial, Spiritual*>—the results of *spiritual activities* in their material embodiment: works of art, literature, socio-cultural events.

Let's consider but one example—that of the design services. The products of design activity are material by form, but have spiritual aesthetic value, which is reflected in the website content and form. Though words of positive evaluation are characterized with high rate of usage in all services proposals, the design sites are notable for high frequency of affective lexemes, in contrast to the rational evaluations of material goods and services. Moreover, the description of previous activity of the designer (put on the "about us" page) usually includes "honours" content block to present the Subject's activity advance with relation to some evaluative scale: getting rewards, taking part in prestigious creative contests.

Product Form: {*Digital, Physical*}. By this parameter digital or virtual products that are delivered electronically (like e-text, graphics, audio and video files, software, financial instruments, etc.) are differentiated from real physical products. The "download" block in the website structure is a typical example.

Environment: {*Virtual, Real*}. The *Virtual* value corresponds to activities that are carried out in the virtual world of the Web or another electronic medium. "Product type" and "environment" combinations differentiate activities by "virtuality extent", e.g., in the sphere of *e-commerce* there are electronic shops that sell physical goods and electronic libraries delivering digital books by download. This distinction has reflections in the website structure: the delivery of physical goods in real environment may cause content blocks that discuss shipping method or present locative and temporal conditions of the activity.

Possible combinations of values of praxeological parameters form feature vectors differentiating activities that underlie websites and corresponding site models. Specific features contribute to the website structure, composition and choice of language means for the constituent pages and their structural blocks.

3. Communicative Parameters

The website considered as a communicative act (CA) addressed by the author to the target audience (TA) is performed in the context of some type of human activity (Ac). Then the author is a Subject, and TA is a Counter-agent of the activity. Information conveyed by means of CA may concern some topic T as well as certain aspects of Ac.

3.1. A site is targeted on realization of ***communicative tasks*** (CT) and corresponding ***reactive tasks*** (RT) for TA.

Communicative tasks

1. Phatic: establish the contact and keep in touch with TA;

2. Illocutionary: convey information about T and/or Ac (in particular, for the purpose of getting information that is essential for Ac);
3. Perlocutionary: to ensure the desirable reaction RT of TA:

Reactive tasks

1. Phatic: TA enters into communication and keeps up the exchange/contact;
2. Cognitive:
   i. TA accepts/learns information about T/Ac;
   ii. TA activates/develops certain attitude to T and/or Ac (with a pragmatic task in mind, the author is seeking for favourable evaluation of Ac by TA);
3. Communicative: in response to CA, TA performs a communicative act that conveys information about TA (knowledge, beliefs, or attitudes) with respect to T or Ac;
4. Pragmatic: TA performs certain non-communicative act Ac', which is presupposed by Ac and embedded in its structure (order/purchase, donation, application for participation, etc.).

Communicative tasks form the hierarchy, in which the performance of the higher-level tasks is presupposed by accomplishment of the lower-level ones. Task structure is determined by ultimate goals and motives, so it may serve a basis for subdivision of websites into three genre groups.

Informative site: CT structure includes RT1—RT2. Site is aimed at modification of beliefs and attitudes of TA by providing information on the topic T, starting from assumption of the general interest of the audience to T. This genre category is somewhat similar to monologue as no correction of the author's knowledge about TA occurs.

Communicative site: CT structure includes RT1—RT3. Site is intended to facilitate human communication and interactions (blogs, forums, social media sites). It provides dialogue or conversation exchange in the course of which mutual beliefs of the participants may be clarified and corrected.

Business site: CT structure includes RT1—RT4. Site is directed toward organization of joint activity. This genre category is somewhat intermediate between informative and communicative ones. The communication is highly stereotyped and the beliefs about the target audience are more or less general (similarly to informative sites). Still, the site structure ensures a feedback for TA on those aspects of the activity that are crucial for effective accomplishment of pragmatic task RT4.

With the view of specifying genre models of websites, it might be useful to combine the obtained genre categorization with a topical one for a unified hierarchy, or a rubricator, of websites to be constructed. Then the feature vector representing specific genre category of, say, the dentist office website would look as follows:

<Business, Institutional Subject, Individual Beneficiary, Services, Unsubstantial, Vital, Physical, Real, Stomatology >

3.2. The instances of different genre categories are structured differently, i.e. the website genre correlates with certain composition(s) of typified structural

constituents: individual pages, sections (groups of pages), and constituent blocks (moves) within the pages.

Genre typing of block/page/section involves functional aspect as well as considerations of communication medium such as interactivity, content dynamics, and modality. Table 1 illustrates this classification in reference to the components of business site. It is significant that many web pages are multi-functional, for example, *Main/ Home page* is intended to welcome TA and inform them of the overall purpose of the site, i.e. to present the Activity with focus on most important aspects (Subject as a site owner, Benefits, Products, News and Events, etc.) and give links to their detailed descriptions (navigation function). Prevailing functionality of the page is resultant of multiple functions of its constituent blocks.

Table 1. Communicative parameters for genre typology of website constituents

Interactivity	
interactive	non-interactive
Forum *Questions and Answers* *Search form* *Shopping cart* *Registration form*	*FAQ(s)*[2] *About us page/section* *Main/Home page* *Contact info* (Subject) *Article*

Dynamics	
invariable content	variable content
About us page/section *Main/Home page* *Contact info* (Subject) *Article*	*Forum* *Questions and Answers* *Commented page* *News*

Modality			
text	image	video	audio
Article *Text portfolio*	*Photo gallery* *Before and After*	*Video gallery*	*Audio library*

Functionality				
Presentational (Ac, Subject)	Informative (Ac, T)	Informative (Product)	Contact	Directive
Main/ Home page *About us page/ section*	*Article* *Gallery* *FAQ* *News* *Staff*	*Catalogue* *Goods info* *Services info* *Price list*	*Contact info* *Registration form*	*Questions and Answers* *Commercial page* *Shopping cart* *Registration form* *Search form*

[2] The *FAQ* page is structured as a succession of question-answer pairs, but unlike *Questions and Answers* pages *FAQs* are non-interactive as they are created by site developers on base of preliminary analysis of possible informational needs of the audience.

4. Genre Schemata and Genre Markers

Website genre corresponds to a number of variants of site structure, which involves the choice of site constituents (sections, pages, and blocks) with their genre types, layout, and hyperlincs. On the page level genre types correspond to ***genre based schematic structures*** composed of functionally determined content blocks.

In case of a textual page, the compositional scheme is described as a succession of text blocks[3], each being a relatively independent and semantically coherent text fragment that represents certain ***content aspect*** of the website pragmatics. Consider the site genre defined as <Business, Services, Institutional>. For this category of sites the presentational *About us* page usually includes the description of the Activity and its Subject and may look like the following succession of content blocks:

<About_Preamble>, <History>, <Advantages>, <Licences>, <About_Coda>.

The proposed formalization of genre schemata of textual pages is based on the earlier works on summarization of scientific papers by the so called "indicator method" [Bljumenau et al. 1981]. According to this method, identification of content aspects is supported by special lexicons of non-topical words used to design and organize scientific text narrative.

The analysis of business websites of Runet has discovered a wide range of characteristic words, distinctive set expressions, verbal clichés that indicate pragmatic content aspects and provide the clues for determining the boundaries of their presentation in the text (text fragments). These words and expressions are specific for particular genres and may be considered as ***genre markers***. For example, the presentational pages of *Services* sites (*About us* or *Home*) usually focus on the "Advantages" aspect that describes favourable features of the proposed services in order to demonstrate the superior position of the activity Subject over the competitors. The corresponding content block could be detected on the page with the help of genre markers "why us", "our advantages", "five reasons" (see table 2 for formal descriptions of Russian markers and counterpart English examples).

We are developing the inventory of Russian genre markers, which are formalized as lexico-grammatical patterns (cf. [Bol'shakova et al. 2006]). The pattern expression may include words, punctuation marks, and slots that are bound with lexico-semantic and grammatical constraints (grammatical class, features, agreement). The pattern elements may be optional (shown in square brackets) and variable (curly brackets). In addition, each pattern is accompanied with text fragment presentation conditions partially based on HTML text formatting tags. They specify the format type (heading, title, list, paragraph, etc.) and position (beginning, end, inside, next to, etc.) of text fragments to be identified (marker fragments) and extracted (aspect fragments). So for any marker fragment identified, say, in the heading, the scope of the aspect fragment may be defined as the list following the marker fragment, or the text following the marker fragment as far as the next marker.

---

[3] In [Kibrik 2009] it is proposed that the linguistic definition of genre and genre categories should be based on the study of types of "passages" (i.e. blocks of genre schemata).

Table 2. Genre markers of the content aspects: "Advantages"

Aspect	Marker pattern	Text presentation		Genre	
		Marker fragment	Aspect fragment	Page type	Site category
Advantages	**"why us"** почему {мы [лучше]; именно мы; именно (*наш* X); Y} ***Lex-sem*** X—<company_type> *компания*; *фирма*; *центр*; *клуб*; *магазин*... Y—<company_name> ***Gramm*** X <Noun, Case=nom, Number=sing> (*наш*, X) <Agree (Gender, Number, Case)>	Head Par_ begin	List_next Text_next	*About us* *Home*	<Business, Services, Institutional>
	English examples: *Why us? Why choose us*; *Why our company?*				
	"our advantages-1" [наши] преимущества	Tit_in	Text	*Advantages* (in *About us* section)	
	English examples: *Advantages; Our Competitive Advantages*				
	"our advantage-2" [в чем] {наши преимущества; преимущества {работы; сотрудничества} с {нами; (*наш* X); Y}} ***Lex-sem*** X—<company_type> *компания*; *фирма*; *центр*; *клуб*; *магазин*... Y—<company_name> ***Gramm*** X—<Noun, Case=instr, Number=sing> (наш, X) <Agree (Gender, Number, Case)>	Head Par_ begin Par _in	List_next Text_next	*About us* *Home*	
	English examples: *Some of the advantages of our company*				
	"five reasons" ([N] *причина*) **[<...>] {{** купить; покупать} [именно] {у нас; в (*наш* X); в Y} ***Lex-sem*** *N*—number/Numeral X—<company_type> *компания; фирма; магазин,..* Y—<company_name> ***Gramm*** X—<Noun, Case=loc, Number=sing> (*наш*, X) <Agree (Gender, Number, Case)> *(N, причина)* <Coord (Gender, Number, Case)>	Head Par_begin Head_in Par _in	List_next Text_ next	*About us* *Home*	< Business, Commerce, Institutional >
	English examples: *Top 5 Reasons to Buy Drect from Knipf*; *A few reasons why you should order from* us; *Top 10 reasons to purchase from Nissan of McKinney*				

5. Conclusion

In the study reported here an attempt has been made to look at the issue of identifying genres on the Web for the purposes of Web generation. The pragmatics based approach to the development of a multifaceted web genre classification is two-level: website genre categories considered with regard to praxeological and communicative context and page genre types specified in terms of functional and medium-related features.

The genre type of a textual page corresponds to a variant of genre schematic structure that is composed of blocks, each representing some content aspect. Genre schemata could be extracted from texts of web pages by using the inventory of genre markers, which is now under development. Genre markers are formally described by common lexico-grammatical patterns, additionally supplied with text presentation (format and position) features. The full-scale repertoire of genre markers described in this way will be useful to not only analyze web pages and identify their genre based schematic structures but also to generate such a structure, expand, and partially populate it with standard content blocks.

References

1. *Bahtin M. M.* (1986), The Problem of Speech Genres [Problema rechevyh zhanrov], in Aesthetics of Verbal Creation [Jestetika slovesnogo tvorchestva], Iskusstvo, Moscow, pp. 250–296.
2. *Bljumenau D. I., Gendina N. I., Dobronravov I. S., Lahuti D. G., Leonov V. P., Fedorov E. B.* (1981), Formalized Summarization by Using Verbal Clichés (markers) [Formalizovannoe referirovanie s ispol'zovaniem slovesnyh klishe (markerov)], Scientific and Technical Information. Series 2. Information Processes and Systems [Nauchno-tehnicheskaja informatsija. Ser. 2. Informatsionnye protsessy i sistemy], № 2 pp. 16–20.
3. *Bol'shakova E. I., Vasil'eva N. E, Morozov S. S.* (2006), Lexicosyntactic Patterns for Automatic Processing of Scientific and technical texts [Leksiko-sintaksicheskie shablony dlja avtomaticheskogo analiza nauchno-tehnicheskih tekstov], Proc. of the 10th National Conference on Artificial Intelligence with International Participation CAI-2006 [Desjataja Nacional'naja konferencija po iskusstvennomu intellektu s mezhdunarodnym uchastiem KII-2006], Vol. 2, Fizmatlit, Moscow, pp. 506–514.
4. *Crowston K., Kwaśnik B., and Rubleske J.* (2010), Problems in the Use-Centered Development of a Taxonomy of Web Genres, in Mehler A., Sharoff S., Santini M. (eds), Genres on the Web. Computational Models and Empirical Studies, Springer, Dordrecht, pp. 69–84.
5. *Kibrik, A. A.* (2009), Mode, genre, and other parameters of discourse classification [Modus, zhanr i drugie parametry klassifikacii diskursov], Problems of Linguistics [Voprosy jazykoznanija], № 2, pp. 3–21.
6. *Mehler A., Sharoff S., Santini M.* (eds) (2010), Genres on the Web. Computational Models and Empirical Studies, Dordrecht: Springer.

7. *Santini M., Mehler A., and Sharoff S.* (2010), Riding the Rough Waves of Genre on the Web, in Mehler A., Sharoff S., Santini M. (eds), Genres on the Web. Computational Models and Empirical Studies, Springer, Dordrecht, pp. 3–31.
8. *Shchipicina L. Ju.* Complex description of genre of computer-mediated communication (on the material of the news agencies web pages) [Kompleksnaja harakteristika zhanra komp'juterno-oposredovannoj kommunikacii (na primere veb-stranic novostnyh agentstv)], available at: http://www.pags.ru/science/conferences/E-Conference/Shipitina.doc
9. *Site Technologies Inc*, available at: http://www.veloxsites.com/
10. *Yandex catalog,* available at: http://yaca.yandex.ru/

PRACTICAL ASPECTS OF LONG-TERM ONTOLOGY-BASED INFORMATION EXTRACTION

Kravchenko A. (anna.kravchenko@interfax.ru),
Pivovarov V. (vasiliy.pivovarov@interfax.ru),
Zharikov A. (alexander.zharikov@interfax.ru)

Interfax, Moscow, Russia

'Ontology-based information extraction' is a subfield of information extraction, where ontologies play an essential role in the process, shaping both system input and target output. There are many different approaches to creating and maintaining an ontology and little work has been done to evaluate and compare the effectiveness of those approaches.

In addition, the practical applications of those systems differ drastically from theory. Architecture that shows good performance in a single test does not necessarily perform as well in the long term.

We conducted an experiment to explore the issues that arise during practical application of OBIE methods and to describe the behavior of ontologies maintained during a long period of time.

In this article we discuss emerging problems and propose working solutions for them as well as the way of evaluation of OBIE systems. Those solutions were successfully implemented in the scan-interfax.ru project and have provided sufficient quality for the commercial use of an advanced entity-based search engine extracting information from news.

Keywords: ontology, information retrieval, search engine, ontology-based information extraction, news extraction, news analysis, ontology population

1. Introduction

The term 'ontology-based information extraction (OBIE) only appeared a few years ago, though some work related to this field has been carried out much earlier. It is a subfield of information extraction, where ontologies play an essential role in the process, shaping both system input and target output.

Information extraction (IE) mostly deals with shallow parsing of the processed data, without attempting a deep linguistic analysis of all aspects of a text. In this way IE systems can be sufficiently fast to deal with the large amounts of web data. At the same time, the text itself may contain conceptual structures and semantic links that are crucial for understanding its meaning and need to be processed thoroughly, especially for domain-specific tasks. Using ontologies allows to combine both those approaches.

Evidently, the quality of the ontology used is critically important for such system to work. There are different ways to create an ontology, a good overview is given in [Wimalasuriya, 2010]. In most cases ontologies are created manually or taken off-the-shelf, some use automatically populated ontologies. Both options have their benefits. For example, a manually created ontology is better for identifying geographical names, while news articles require constant influx/addition of new data due to world dynamics.

It also should be noted that practical applications of those systems differ drastically from theory. Architecture that shows good performance in a single test does not necessarily perform in a satisfactory manner in the long term. New entities that are absent in fixed/manually created ontologies appear constantly in the world, and for automatically updated ontologies errors tend to accumulate.

We've conducted an experiment to explore the issues that arise during practical application of OBIE methods and to describe the behavior of an ontology maintained during a long period of time.

In this article we discuss issues that emerged during the experiment and propose working solutions for them, as well as a way of evaluating OBIE systems.

2. Experiment details and system architecture

Most rule-based systems (except for [Hwang, 1999]) use manually constructed ontologies which are not updated. Updating ontology automatically increases recall and also provides opportunity for research.

For our experiment we used the scan.interfax.ru system, which is focused on news analysis and entity-oriented search. The demo version is free, though available with limited functionality.

The system is mostly rule-based to maintain precision, although it uses some statistical algorithms such as Bayesian and SVM classifiers. Ontology classes and relations (Tbox) are set up manually while the collection of entities (Abox) is constructed and updated automatically from news articles texts using a bootstrapping approach. Scheme of the system's architecture can be seen on Fig. 1.

The procedure of adding entities to the database consists of two stages.

In the first stage entities are extracted from the article. It is a part of a more general procedure, which also reveals morphological and syntactic structures, conducts anaphora resolution, extracts keywords and key sentences, etc. The key feature of this entity extraction approach is its independence from any time-dependent object databases. In other words, the system aiming to extract person or organization entities does not use any lists of real world's ones and only operates with semantic and morphological dictionary data.

At second stage extracted essences are identified with an entity stored in the database. If such an entity is not found, a new one is created. The stored entity should represent a real world object and has links to any information about it.

A detailed description of the algorithm can be found in [Zharikov, 2011].

For our experiment we chose a time interval from 01.01.2013 to 31.12.2013 containing approximately 2,500,000 documents from the news stream. The identification

of entities was held with only rule-based precise algorithms used (all statistical procedures were switched off). At its initial state the database was empty. At the end of the procedure there were approximately 1.2 million of entities (persons and organizations) collected.

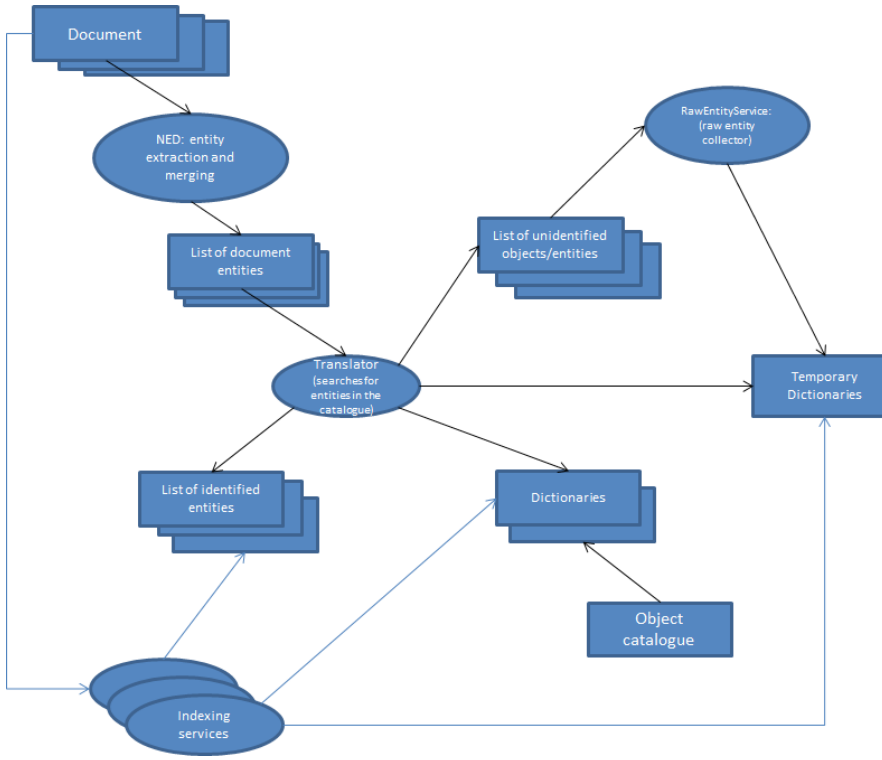


Fig. 1. Process of entity extraction in Scan system

Three values were measured: number of precise entity identifications, number of unresolved ambiguities and number of identification missing (candidates for identification were not found in the database).

We also measured the lifespan of extracted entities and their number of occurrences.

3. Discovered issues

In this article we will analyze the extraction of person entities since they are more illustrative. In context of analyzed issues organization extraction specifics is very similar.

3.1. World dynamics problem

World dynamics is the main source of ambiguity. It is also unavoidable.

New names appear in the news all the time. Also some roles shift from one person to another, for example, “The British Prime Minister” may refer to different people during different time periods. A new person with the same name may appear or the person could change his or her role. All this leads to ambiguity of interpretation.

Ideally, date range for each entity should be stored in the database, and the list of relevant entities should be generated accordingly to the date assigned to the processed document. The database should also be updated timely.

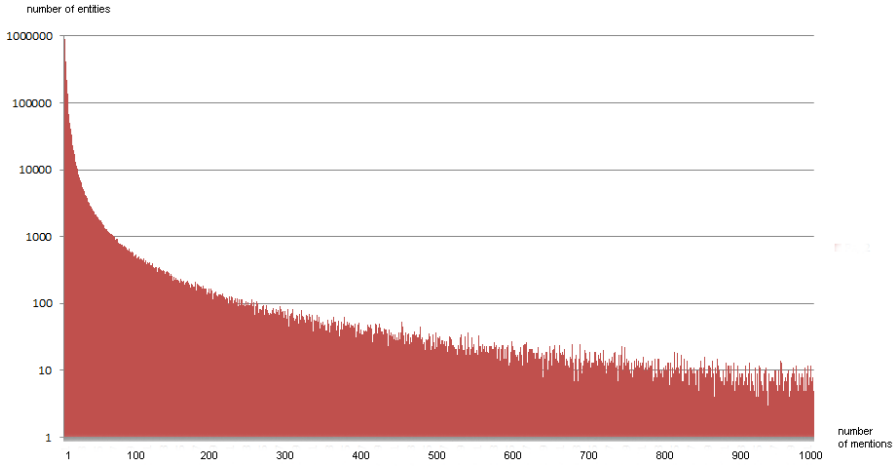


Fig. 2. Number of entities corresponding to number of mentions

As we can see on Fig. 2, most of the entities are mentioned only a small number of times or even once. Even further, Fig. 3 shows that the lifespan of most entities doesn’t exceed one day.

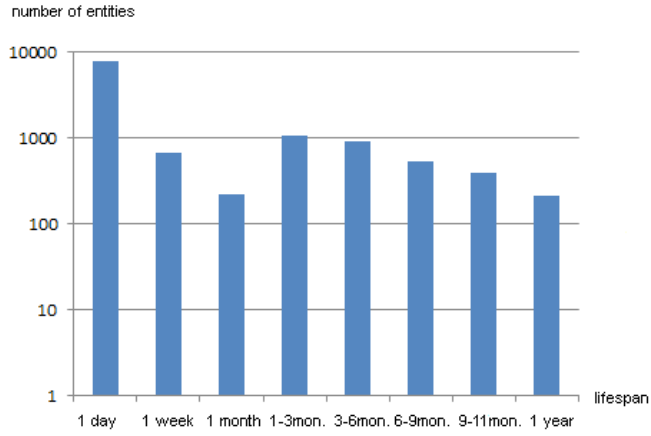


Fig. 3. Number of entities corresponding to their lifespan (we removed entities that were mentioned less than 20 times from this chart)

We should also note that every entity belongs to one of the three following types.

1. Rare entities
   Number of mentions is very small (less than 10 mentions usually). A good example would be a school headmaster that sometimes appears in the regional press. Lifespan is fairly long.
   Those entities usually come along with proper qualifiers and are easy to extract correctly.
   Qualifiers (not to be confused with qualifiers in formal semantics) are noun groups that describe person's role. For example, in “the great painter Van Gogh” the phrase “the great painter” would be a qualifier.

2. Constant objects.
   Those fall into one of two subcategories:
   a. Contemporary public figures (mostly politicians), those have high occurrence and long lifespan.
   b. Historical figures, are characterized by low/medium occurrence and very long lifespan, usually exceeding system’s lifespan.

   It is hard to extract roles for those persons, because those roles are considered common knowledge and a mention of such a person usually comes without a proper qualifier.
   However, using an ontology allows to extract such entities successfully, since in a large enough corpus it is eventually possible to find a good qualifier. Nevertheless it is recommended to have a small list/dictionary of historical persons, because some of them may be rather rare.

3. Cluster objects
   Objects of this type have high occurrence and short lifespan. They are usually connected to the same news story.
   They mostly appear with roles and are easy to. To enhance recall it is also recommended to use topic detection methods.
   Cluster object are the main reason to update database timely.

A similar classification can be proposed for organization entities:

1. Rare or common entities such as local stores or motorcar factories. They usually have long lifespan, small number of mentions and the reader is not supposed to know about them, so they are always mentioned with proper qualifiers and are easy to extract.
2. Constant objects that readers are supposed to know about. They tend to have high occurrence, long lifespan and no qualifiers. Oil companies can be a good example of these. This type consists mostly of present-day companies, since in comparison to historical person objects historically-significant organizations are extremely rare.
3. Cluster objects are exactly the same as person cluster objects.

Relying on this classification, we can conclude that updating the database regularly is important. Editing database manually requires a lot of human effort. Automatic updating seems to be the solution, however it causes two other problems.

3.2. Long tails

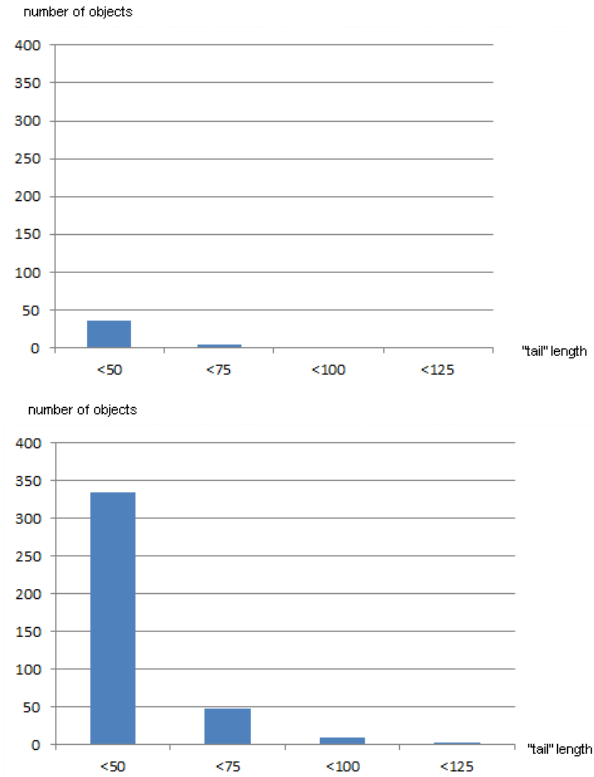


Fig. 4. Number of long tails after 3 months and after 6 months of documents

The entity extraction procedure can fail to connect person name with a role or to extract a full name. It can also fail at identifying the entity if it is misnamed. Merging similar names by roles helps, but the same name can correspond to different roles. Persons with the same name and different roles are also considered different entities and even simple roles are difficult to merge, so this creates even more problems.

In automatically updated ontology this leads lead to a long “tail” of entities: for example the tail for Muammar Gaddafi can include a succession of pair derived from (Muammar, Mummar, Muamar…)×(Caddafy, Kaddafy, …). Tail for Jim Jarmush can contain such roles as “acclaimed director and musician” and “the creator of the film ‘Limits of control’”.

It influences both precision and recall rate (see Fig. 5) and can potentially slow down the system.

This problem can be fixed by a name merging algorithm, as we will show later.

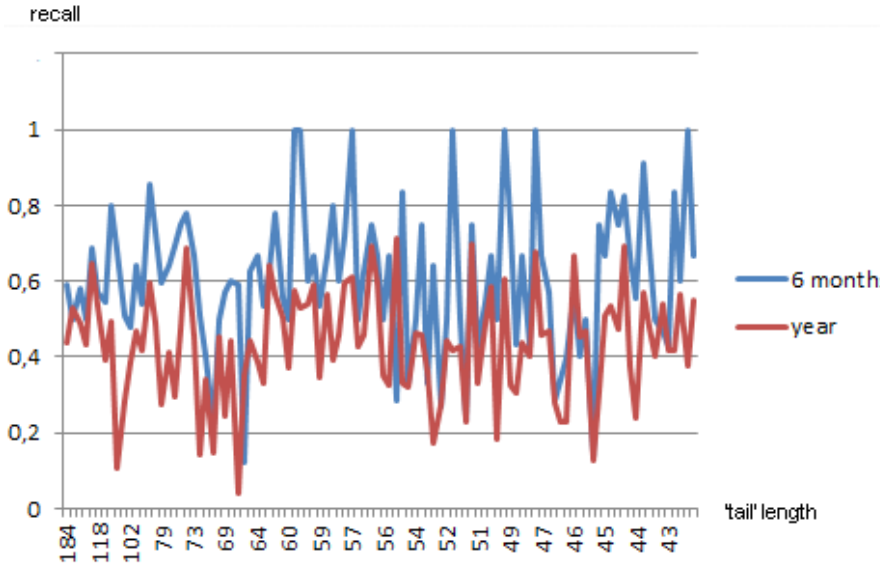


Fig. 5. Recall difference for entities with long "tails" after 6 month and a year of database populationg, shown on the x axis is the tail's length. we can see recall values for long-tail entities computed for their mentions in July and for mentions in December, where recall = the number of identified entities divided by the number of matching strings (both found and missing/unresolved entities).

3.3. Error accumulation

Existing entity extraction algorithms are not 100% accurate and tend to assign incorrect roles to entities. Even if we obtain an algorithm with 100% precision, the information itself can be unreliable, for example, the author may assign a wrong role to a person. There is also the factor of words having multiple meanings. If you take a geographical dictionary containing several hundred thousands of entries and tag a random text with it, few words wouldn't be considered a geographical object for one reason or another. A similar phenomenon occurs when using a sufficiently large list of organizations. Conflicts arise both within dictionary and with objects of other types.

In practice this inaccuracy becomes crucial. A procedure with 99% of precise identifications creates the impression of being almost perfect, but in the long term it leads to a burst of identification errors, because the ontology is being filled with incorrect entities.

For automatically updated ontologies we consider this problem the most serious.

4. Proposed solutions

1. The first and the foremost necessity for a practically applied OBIE system is to update the ontology regularly. The best way is to use an automatically updated ontology. It will help dealing with the influx of new entities in the news stream. However, this may lead to ontology degradation.

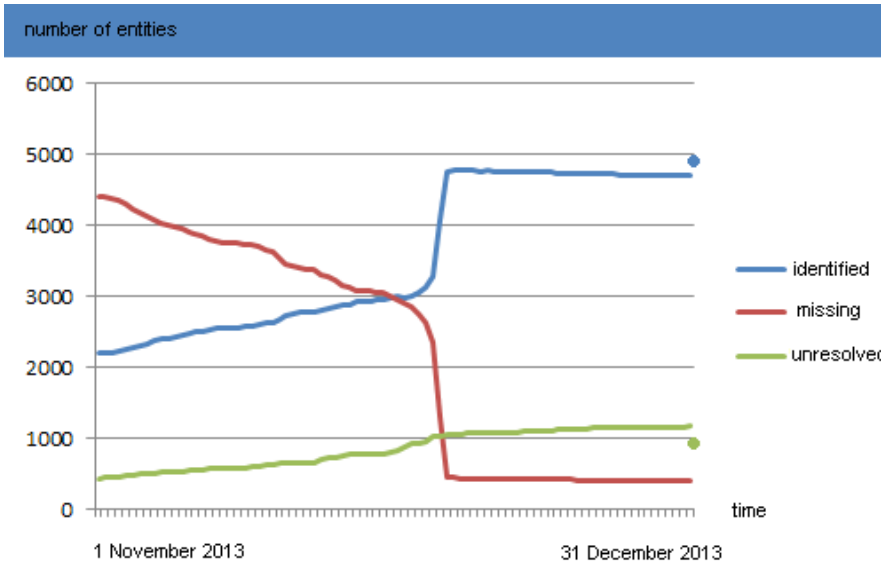


Fig. 6. Number of unresolved, missing and successfully identified entities

To evaluate the degradation rate for the object identification system an experiment was performed. Namely, the fixed set of documents was passed through entity identification system under the different states of entity catalog, which corresponds to the increasing values of date. In the certain case considered the document set was a November subset of the general document stream. This circumstance resulted in abrupt grow in the middle of the Fig.6 which corresponds to November period of ontology population procedure

We can see that the number of precise identifications is almost constant and begins to fall very slowly from the middle, while the missing and unresolved rates are inversely related (the growth is the faster, the nearer the database filling procedure approaches the testing interval). We can also see that the growth of unresolved entities rate is nearly linear.

The speed of the identification degradation due to ambiguity growth (the slope angle of the green line on the graph) also seems to be a good measure of quality, which can be used to test overall system quality and internal consistency and perhaps to compare different systems between each other.

2. Merging of long tails can have significant effect on recall rate.
   The dot on the right end of Fig.6 refers to the number of resolved entities after applying a merging algorithm, based on editing distance and entities' connection to the same organization (people with similar names belonging to the same organization were considered the same person).
3. To deal with the accumulating errors we propose using separate databases for each time period (for example, 6 months).

Important entities are repeated often while mistakes are rarer and filling a new database allows to "clean up" the system from errors and outdated entities.

There is a chance of losing an important entity, but Fig. 6 demonstrates that most frequent entities (the list of entities was chosen by assessor) all have successfully migrated from the older database to a newer one. Entities that appear in the stream from time to time with a sufficiently large interval (months) may still be lost, but this negative effect is not essential for most applications. It can also be compensated for by having a small dictionary of "main" entities, mostly of type 2 (section 3.1) (and by manual database correction).

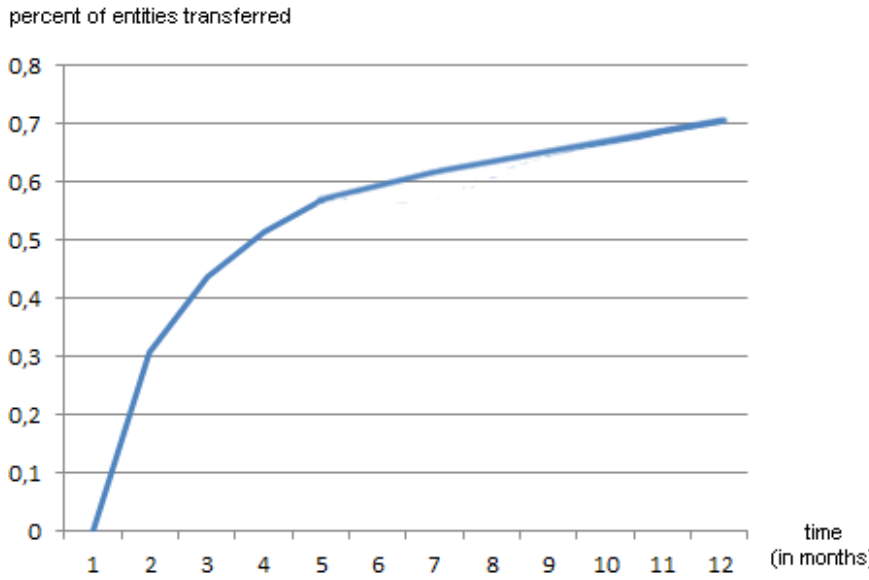


Fig. 7. Percent of entities transferred to the new database during the year

5. Results and discussion

This logic was successfully implemented in the scan-interfax.ru project and has provided sufficient quality for the commercial use of advanced entity-based search engine.

Performance was analyzed by an assessor. For companies the system achieves 99% precision and 78% recall for raw entities and 100% precision and 95% recall for "active" entities, verified by a human (note, that we do not consider entities splitting in long tails an error), and for persons precision is 90% and 90% recall for raw entities and 95% precision and 97% recall for "active" entities.

6. Conclusion

We have described the main aspect of OBIE system behavior in the long term and proposed some approaches to solving emerging problems, which allow us to fill a database with entities from a news stream. Precision provided by algorithms approaches 100% and recall is high enough to set a standart for training a statistical identifier or to make various conclusions automatically. The recall rate is influenced much by the ambiguity caused by world dynamics and by the sensitivity of the system to mistakes in the entity extraction procedure and in the semantic comparison. Nevertheless the negative influence is not crucial for small time intervals (<2 years). The novel approach is proposed to account for these effects by means of using a time dependent database. The negative side of that approach is a loss of the possibility to bring together entities which appear in the stream from time to time with a sufficiently large interval (months). This negative effect is not crucial for most applications. The speed of the identification degradation due to ambiguity growth seems to be a good measure of quality, which can be used to test overall system quality and internal consistency and perhaps to compare different system between each other. External merge procedures can be implemented periodically to prevent degradation effect

The logic of the precise identification was successfully implemented in the scan-interfax.ru project and has provided sufficient quality for the commercial use of an advanced entity-based search engine. The implementation of supervised methods based on automatically collected standart datasets is in the development stage. The core logic proposed is suitable for a multilingual system though depends much on the entity extraction procedure and entity ontological interpretation which are often language specific. The core logic shown on the example of person entities seems to be suitable for other entity types. The same principles are used in the Scan project to identify organizations and to fill the database with the missing ones (the analog of roles are organization types and locations). The identification of natural language named entities of arbitrary types (such as films, car models, brands at whole, animal names etc.) has been just implemented and now is passing the testing stage.

It includes automatic filling of not only unknown object database but also a base of object types and its semantic hierarchy. The main disadvantages of the proposed approach include lack of universality for different content types and extreme sensitivity to extraction and comparison rules incorrectness.

References

1. *Zharikov A., Kristalovsky K., Pivovarov V.* Information Retrieval System for News Articles in Russian. Web of Data: The joint RuSSIR/EDBT 2011 Summer School, pp. 5–14 (2011).
2. *Dredze, M., McNamee, P., Rao, D., Gerber, A., Finin, T.* Entity disambiguation for Knowledge Base Population. In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 277–285. Beijing, China (2010)
3. *Anastácio, I., Martins, B., Calado P.* Supervised Learning for Linking Named Entities to Wikipedia Pages. In: Proceedings of the Text Analysis Conference, Nov. 2011
4. *Bunescu, R. and Pasca, M.* Using encyclopedic knowledge for named entity disambiguation. In Proceedings of the European Conference of the Association for Computational Linguistic, EACL '06. (2006)
5. *Artiles J., Borthwick A., Gonzalo J.* WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Tasks. In Third Web People Search Evaluation Forum (WePS-3), CLEF 2010.
6. *Hien T. Nguyen, Tru H. Cao.* Named Entity Disambiguation: A Hybrid Statistical and Rule-Based Incremental Approach. The Semantic Web. Lecture Notes in Computer Sci-ence Volume 5367, 2008, pp. 420–433
7. *Davis A., Veloso A., Soares da Silva A.* et al..: Named Entity Disambiguation in Streaming Data. ACL The Association for Computer Linguistics, pp. 815–824 (2012).
8. *Hassell J., Aleman-Meza B., Budak Arpinar I.* Ontology-driven automatic entity disambiguation in unstructured text. ISWC'06 Proceedings of the 5th international conference on The Semantic Web,
9. *Daya C. Wimalasuriya, Dejing Dou.* Ontology-based information extraction: An introduction and a survey of current approaches. J. Information Science 36(3): 306–323 (2010) [http://www.informatik.uni-trier.de/~ley/db/journals/jis/jis36.html#WimalasuriyaD10]
10. *Chung Hee Hwang.* Incompletely and Imprecisely Speaking: Using Dynamic Ontologies for Representing and Retrieving Information. KRDB, volume 21 of CEUR Workshop Proceedings, page 14–20. CEUR-WS.org, (1999)

ТЕЛО В ДИАЛОГЕ: ОРИЕНТАЦИЯ СОМАТИЧЕСКИХ ОБЪЕКТОВ И ВЫРАЖЕНИЕ ОТНОШЕНИЙ МЕЖДУ ЛЮДЬМИ

Крейдлин Г. Е. (gekr@iitp.ru),
Переверзева С. И. (P_Sveta@hotmail.com)

Российский государственный гуманитарный университет, Москва, Россия

Телесные, или соматические, объекты, их признаки, действия с ними и над ними были предметом исследования целого ряда работ. Однако в этих работах мало внимания уделялось связям физических компонентов, относящихся к телу человека или к его частям, и психологических характеристик человека. Известно, например, что чрезмерно высокий или низкий рост человека, излишние толщина или худоба его тела способствуют возникновению комплексов, вплоть до комплекса неполноценности, а потому телесные аномалии многим людям мешают вступать в коммуникацию.

В настоящей статье нам бы хотелось несколько продвинуться на пути к решению задачи системного описания соответствий физических и психических свойств человека. С этой целью на примере анализа признака «ориентация соматического объекта» мы хотим показать его связь с психологическими характеристиками человека — прежде всего, теми, которые говорят об отношении человека к собеседнику в актуальном диалоге.

Ключевые слова: тело, соматический объект, пространственная ориентация, диалог, психологические характеристики, отношения между людьми

HUMAN BODY IN A DIALOG: THE ORIENTATION OF SOMATIC OBJECTS IN ITS CONNECTION WITH HUMAN RELATIONS

Kreydlin G. E. (gekr@iitp.ru),
Pereverzeva S. I. (P_Sveta@hotmail.com)

Russian State University for the Humanities, Moscow, Russia

The main objective of the paper is to examine relations between corporeal, or somatic objects and some psychological aspects of human behavior, namely the relations between the communicators in a dialog.
Somatic objects have been investigated from different points of view. Mostly, linguists and specialists in nonverbal semiotics have described names, features and significant actions performed by or with different somatic objects, virtually leaving aside sign manifestations of correlations between physical (corporeal) and psychological (ethical, aesthetical, etc.) aspects of human behavior. It is well-known that if a man is lengthy or extremely short or if he is too fat or scrawny he feels bad about his deficiency. Also, it is known that many corporal defects impede or aggravate proper communication. Here we undertake a few preliminary steps in solving the problem of the systematic description of the correlations between physical and psychological properties of humans. We consider one corporal feature — "spatial orientation" — that many body parts possess and describe its relations with the psychological characteristics of interlocutors. The explication of the notion "orientation of somatic object" is given and two Russian linguistic representations of spatial orientation are discussed. The linear representation corresponds to the linguistic construction *X V Y-Instr Prep Z*, where X is an oriented object, V is a verb of orientation (e.g. *smotret' na chto-libo* 'to face smth. (e.g. about the buildings)', *byt' napravlennym na chto-libo* 'to be directed at smth.'), Y is a <so called> salient part of the object X, Prep is a preposition and Z is an orienting point. The angular representation accords with the construction *X V Prep Z pod uglom* 'at an angle' *Q* (where Q is a degree of the angle).
The basic part of the paper is devoted to the correspondence between the corporal orientations which are computed by these representations and which are expressed either verbally or nonverbally (based on the Russian body language) and ethical features of humans participating in an actual dialog. Thus, different types of bows conform regularly to the features 'respect to the addressee', 'veneration of the addressee' or just 'warm feeling to him / her'.

Key words: body, somatic object, spatial orientation, dialog, psychological characteristics, relationships between humans.

1. Введение. Постановка задачи

Ориентация соматического объекта является одним из важнейших признаков, описывающих как самого человека, так и разные аспекты его

коммуникативного поведения. Под **ориентацией объекта** (в том числе соматического) мы будем в первом приближении понимать его направленность на некоторый объект (причём не обязательно телесный) или на часть пространства, ср. такие языковые обозначения ориентации, как *показать пальцем на здание* или *лежать головой на север*. Для дальнейшего уточнения понятия пространственной ориентации требуется ввести фигуру наблюдателя. Рассмотрим предложение *Дом ориентирован балконами на речку / на восток*. На наш взгляд, его семантическая структура представляется следующим образом: ‘если вообразить себе, что на балконе стоит человек и смотрит вперёд, то речка мыслится как тот объект, на который направлен его взгляд (соответственно, восток мыслится как та часть пространства, на которую направлен его взгляд)’.

Основное внимание мы уделим признаку «ориентация соматического объекта» и тем его значениям (values), которые выражают отношения между людьми — участниками актуального диалога. Сразу же отметим, что выражения как значений данного телесного признака, так и отношений между людьми могут быть вербальными или невербальными. Тем самым, в работе решается проблема соответствия отдельных значений телесных и психологических характеристик человека, и эта проблемы является составной частью более широкой проблемы мультимодальности. Последняя заключается в описании важнейших способов взаимодействия в коммуникативном акте вербальных и невербальных знаков. В качестве семиотических кодов мы далее рассматриваем русский язык и русский язык жестов[1].

2. Основные понятия, используемые в работе. Линейная и угловая ориентация

В диссертации [Переверзева 2013], посвящённой детальному анализу признака «ориентация» (в частности, «ориентация соматического объекта») и его значений, были описаны два основных способа знаковой репрезентации ориентации, получивших названия **линейной** и **угловой ориентаций**.

В русском языке для выражения линейной ориентации объектов широко используется конструкция вида X V Y-ом Prep Z. Здесь X, Y, Prep, Z и V — это переменные, на места которых подставляются имена ориентируемого объекта (X), его выделенной части (Y), предлога (Prep), названия ориентира (Z) и обозначения действия или ситуации (V), в которой объект X ориентируется относительно Z.[2]

Предложения, построенные в соответствии с этой конструкцией, это (1) *Церковь ориентирована алтарём на восток*; (2) *Дом выходит окнами в сад*;

---

[1] Поскольку основу языка жестов составляют знаковые движения и положения тела или его частей, неудивительно, что вместо термина *язык жестов* в литературе часто используется термин *язык тела*. Мы тоже позволим себе им пользоваться.

[2] Данная конструкция была предметом внимания в работах [Апресян 2008], [Подлеская, Рахилина 2000] и [Рахилина 2000].

(3) *Воланд протянул руку ладонью кверху* и (4) *Он лежал головой на север, а лицом на запад.* На место переменной X в приведённых фразах подставляются имена ориентируемых объектов — *церковь*, *дом*, *рука* и *он*. Переменную Y замещают имена частей ориентируемых объектов, но не любых, а только тех, положение которых существенно для определения ориентации самих объектов. Такие части ориентируемых объектов получили название **выделенных** (здесь это *алтарь*, *окна*, *ладонь*, *голова* и *лицо*). Переменная Z — **ориентир** — замещается в приводимых предложениях именами *восток*, *сад*, *верх*, *север* и *запад*, а позицию Prep заполняют предлоги *на* и *в*. Кроме того, на место сочетаний Prep Z разрешается также подставлять пространственные наречия — обычно в тех случаях, когда ориентиром является какая-то часть пространства, ср. наречие *кверху* в примере (3). Наконец, переменную V замещают предикатные единицы *быть ориентированным*, *выходить*, *стоять*, *лежать*, *смотреть на*, *повернуть* и др.

Данная ориентация получила название *линейной* потому, что конструкция X V Y-ом Prep Z передаёт идею указания на более близкое расположение **выделенной части** относительно ориентира по сравнению с другими частями ориентируемого объекта, а смысл ‘близости’ удобнее всего изображать **геометрически при помощи отрезка**, соединяющего две точки — ориентир Z и выделенную часть Y объекта X.

Второй способ знаковой репрезентации пространственной ориентации телесных объектов называется **угловым представлением ориентации**. Отметим, что если в вербальном отражении линейной пространственной ориентации основным элементом являются глаголы и иные предикатные единицы, то в вербальном представлении угловой ориентации основным элементом являются наречные образования и предложно-падежные формы, а также отдельные редкие существительные, ср. слова *вполоборота*, *косо*, *искоса, под углом*, *под наклоном, профиль* и др.

В формальном описании угловой ориентации, которое мы даём здесь с некоторыми незначительными изменениями по сравнению с диссертацией [Переверзева 2013], существенно используется понятие **ориентируемого объекта правильной формы**.

Пусть дан физический объект, обладающий следующими свойствами: (1) внутри него можно представить декартову систему координат, состоящую из перпендикулярных друг другу осей, причём (2) измерения длины, ширины и других пространственных параметров данного объекта производятся по одной из этих осей. О таком объекте мы говорим, что он имеет правильную форму, и называем его *объектом правильной формы*. Оси внутри объекта правильной формы (в общем случае их три — вертикальная, горизонтальная и сагиттальная) в тех ситуациях, когда данный объект является ориентируемым, мы называем **осями ориентируемого объекта**.

При угловой репрезентации телесной ориентации указывается, что одна из осей ориентируемого объекта правильной формы (объекта X) образует угол с воображаемой прямой, соединяющей этот объект с ориентиром Z (далее такую прямую мы будем называть кратко *прямая “объект — ориентир”*). Иначе

говоря, угловая репрезентация телесной ориентации задаётся, прежде всего, **углом ориентации**, под которым понимается угол между одной из осей ориентируемого объекта правильной формы в некотором его положении и прямой «объект — ориентир».

Угловому представлению ориентации отвечают, по меньшей мере, две языковые конструкции — X V *под углом* Q Prep Z и X V Y-ом *под углом* Q Prep Z (место переменной Q заполняется числом, обозначающим величину угла ориентации). Конструкции X V под углом Q Prep Z соответствует предложение (5) *В глаз человека, стоящего далеко от озера, попадают солнечные лучи* (X)*, отбрасываемые* (V) *водной поверхностью под небольшим* (Q) *углом к* (Prep) *ней* (Z), а конструкции X V Y-ом *под углом* Q Prep Z — предложение (6) *Вася* (X) *стоит* (V) *спиной* (Y) *к* (Prep) *фотокамере* (Z) *под углом градусов в 45* (Q). Отметим, что некоторые из переменных, входящих в конструкции угловой репрезентации, в реальных предложениях могут оставаться незамещёнными, а вместо сочетания *под углом Q* могут выступать наречные образования. Например, предложению *Петя стоит, повернувшись к Васе боком* соответствует следующее представление: сагиттальная ось тела Пети расположена примерно под прямым углом относительно тела Васи. Здесь в роли X выступает *Петя* (‘тело Пети’), в роли V — предикатная единица *стоит, повернувшись*, в роли Prep Z — предложно-падежная группа *к Васе*, а вместо единицы *под углом Q* — наречие *боком*. Выделенная часть Y — это ‘лицо Пети’, но в данном предложении переменная Y остаётся незамещённой.

Линейное и угловое представления ориентации телесных объектов поразному отражаются в описании языковых единиц. Так, существуют ситуации, когда один из двух способов представления ориентации предпочтительнее другого или вообще является единственно возможным. Примером ситуации, когда существует идиоматичный способ обозначить линейную, но не угловую ориентацию, является ситуация контакта выделенной части ориентируемого объекта с ориентиром. В этом случае построение угла между осью ориентируемого объекта и прямой «объект — ориентир» затруднено. Такую ситуацию иллюстрирует, например, предложение (7) *Он уткнулся лицом в подушку*, где речь идёт об ориентации головы. Говорить об угловой ориентации головы по отношению к подушке тут было бы крайне странно, в то время как линейный способ представления данной ориентации не только возможен, но и естественен. Ориентируемым объектом здесь является голова (имя объекта здесь опущено), выделенной частью головы — лицо, ориентиром — подушка, а значением признака линейной ориентации оказывается «близость лица к подушке».

Предложение (8) *Вытирала руки о фартук и усаживалась <…>* ***полубоком ко мне, вполоборота к маме*** является примером языковой реализации противоположной ситуации, а именно когда описание на языке угловой ориентации допустимо и предпочтительно, а описание на языке линейной ориентации неудачно. В приведённой фразе имеются слова *полубоком* и *вполоборота*, и оба с достаточной степенью определённости задают ориентацию тела относительно объектов, обозначаемых, соответственно, словами *мне* и *маме*. Между тем данную ориентацию невозможно выразить только при помощи

конструкции X V Y-ом к Z (где имя X опущено, V — *усаживалась*, Y — ‘лицо’, а Z — *я / мама*), не потеряв при этом важную смысловую информацию о положении тела, которую заключают в себе слова *полубоком* и *вполоборота*.

Ситуация, обозначаемая воинской командой *Кругом!*, допускает, формально говоря, оба вида репрезентации. Угол поворота человека, подчиняющегося этой команде, — 180°, и этому углу в точности соответствует замена выделенной части ориентируемого объекта на противоположную ей часть того же объекта. Иными словами, до исполнения команды человек стоял лицом в некоторую сторону, а после её исполнения в ту же сторону стала направлена его спина. Между тем геометрия круга и углы поворота, которые широко используются в планиметрии, делают более предпочтительной трактовку наречия *кругом* именно как реализацию поворота на угол 360°. Такая интерпретация изменения ориентации следует из семантики самого слова *кругом*.

3. Представления пространственной ориентации соматических объектов и отношения между людьми

Теперь, после того как был введён необходимый концептуальный и терминологический аппарат ориентации, мы покажем, каким образом признак ориентации соматического объекта и два представления пространственной ориентации связаны с разными знаковыми способами выражения смыслов, касающихся отношений между участниками данного коммуникативного акта.

Когда речь идёт о пространственной ориентации человека, основными, очевидно, являются три направления, или три ориентации, а именно вертикальная (вверх — вниз) и две горизонтальные (влево — вправо и вперёд — назад). Со словами, в семантике которых содержится указание на эти ориентации, связаны разного рода оценки поведения одних людей по отношению к другим, прежде всего этические оценки.

Рассмотрим такие единицы, как *подниматься в чьих-то глазах* или *опускаться* (в переносном значении). С ними связаны оценки морального одобрения и, соответственно, осуждения человека или его поступка. Движение вверх — это движение восходящее, это движение к Небу (ср. глаголы *подниматься*, *возвышаться*, сочетания *во весь рост*, *воздевать руки к небу*), а движение вниз связано с силой тяготения.[3] Ср. такие слова, как глагол *пасть* и существительные *дно*, *низ*, *низость*: (9) *Низость — само слово указывает на тяготение, на феномен силы тяготения*; (10) *Благородное действие может принизить, если нет необходимой силы того же уровня*. *Низкое и поверхностное находятся на одном уровне*. Не случайно *кланяться*, обозначающее движения головы и корпуса вниз, часто воспринимается как *унижаться*, ср. (11) *Я не стану тебе кланяться!*.

---

[3] Семантика движения по разным пространственным осям давно обсуждается в литературе, в частности, в работах, посвящённых метафоре, см., например, Lakoff, Johnson 1980.

Если направление движения и ориентация по вертикальной оси связаны с движениями вверх (к Богу, к Небу) и вниз (от Бога, в Преисподнюю), то направление движения по горизонтальным осям соотносится с движением к людям. На это указывают отдельные глаголы воображаемого или реального движения (*приблизиться*, *соединиться в одно*, *слиться*, *столкнуться*), глаголы, обозначающие влечение одного человека к другому (*Меня к нему тянет / влечёт / притягивает*), а также глаголы, выражающие человеческие отношения или их изменение, типа *разорвать* или *сломать* (*отношения*). Все приведённые единицы говорят о том, что отношения между людьми изменились: стали либо лучше, либо хуже. Иными словами, переносные значения пространственных глаголов ориентации по горизонтали свидетельствует об изменении человеческого измерения.

Обсудим теперь некоторые невербальные единицы из семантического класса **поклонов**. Хотя они уже были подробно разобраны в монографии [Крейдлин 2002] и работе [Морозова 2006], некоторые аспекты остались в этих публикациях ещё не до конца разобранными. Именно на них мы остановимся.

Напомним, что формально виды поклонов различаются разными углами наклона корпуса жестикулирующего к адресатам самых разных типов — к людям, изображениям сакральных объектов или каким-то иным ориентирам. Было показано, что для описания поклонов более приспособлен язык угловой ориентации, поскольку величина угла ориентации при исполнении поклона данного вида зависит от типа адресата и отношения к нему жестикулирующего.[4]

Типовыми русскими глаголами, за которым стоят поклоны разных типов, являются глагол *поклониться* и — в редких случаях — глагол *наклониться*. Отметим, что *наклониться* чаще выражает незнаковые действия (ср. *наклониться к кому-то, перед кем-то* или *наклониться* при исполнении какого-то упражнения в утренней гимнастике; в первом случае более чем одноместный предикат *наклониться* — это глагол ориентации, а во втором случае это одноместный предикат местоположения тела из семантического класса глаголов местоположения).

Разберём подробнее пространственные параметры угловой ориентации тела в ситуации некоторых русских поклонов. Общая закономерность здесь такова: чем глубже позитивные чувства жестикулирующего к адресату, тем больше угол наклона тела в поклоне. Если отношение жестикулирующего к адресату чисто формальное, то даже при выражении позитивного чувства угол наклона невелик (ср. разного рода **кивки** или **академические поклоны**, ср. также фразу *В знак уважения к позиции капитана Арсений наклонил голову* (Е. Водолазкин)). Отметим, что тело жестикулирующего во всех русских поклонах ориентировано передней частью в сторону адресата, а глаза — не в случае религиозных поклонов в православных храмах — тоже смотрят на адресата. Перед сакральными объектами и во время исповеди жестикулирующий смотрит вниз. Этим выражается отношение глубокого почтения и смирения.

---

[4] [Переверзева 2013].

Вообще, отношения смирения, стыдливости, скромности, то есть отношения, в которых жестикулирующий ставит себя ниже адресата, выражаются склонённой головой и опущенными глазами, ср. такие выражения, как *<стыдливо> потупиться*, *<скромно> опустить глаза*, *<смиренно> склонить голову* (эти единицы разбирались, в частности, в работе [Крылова 2010]). Между тем стыдливость другого рода, а именно смешанная с заинтересованностью жестикулирующего в адресате и вниманием к нему, часто выражается следующим **жестовым комплексом**: «голова жестикулирующего опущена вниз, а глаза время от времени смотрят (*косятся*) на адресата», ср. строки из известной песни «Подмосковные вечера» (12) *Что ж ты, милая, смотришь искоса, // Низко голову наклоня?*.

Таким образом, пространственный аспект семантики наречия *искоса* хорошо передаётся на языке угловой ориентации. Физическая реализация жеста **смотреть искоса <на кого-л.>** складывается из трёх компонентов. Один из них — это действие, выполняемое глазами, а именно, как мы только что сказали, глаза жестикулирующего смотрят на адресата, а два других — ориентации глаз и головы. Как известно, в устном диалоге лицо (передняя часть головы) и глаза жестикулирующего в норме направлены на собеседника. Тут, однако, лицо жестикулирующего направлено не в ту же сторону, что глаза, в результате чего образуется угол между прямыми, соответствующими ориентациям головы и глаз.

В диссертации [Переверзева 2013] было показано, что несовпадение ориентации головы и глаз и наличие между ними относительно небольшого угла объясняет возникновение переносного значения у наречия *искоса*. Действительно, во многих толковых словарях для слова *искоса* выделяются, помимо исходного, переносные значения. Например, в словарях [Ожегов 1983] и [Кузнецов 1998] слово *искоса* в исходном значении толкуется как «не прямо, скосив глаза», а в переносном — как «недоброжелательно, с подозрительностью [смотреть. — Г.К., С.П.]». Переносное значение у слова *искоса* возникает, на наш взгляд, не случайно, но для объяснения его появления нужно выйти за пределы собственно русского языка и обратиться к русскому языку жестов.

Ранее было показано ([Крейдлин 2012]), что физическая реализация жестов во многих случаях является композициональной не только по форме, но и по смыслу: смысл многих жестов складывается из семантических компонентов, соответствующих отдельным компонентам его физической реализации. Иначе говоря, такие аспекты телесной ориентации, как (а) близость / удалённость выделенной части ориентируемого объекта к ориентиру / от ориентира и (б) угол между осью соматического объекта при его обычной ориентации и той же осью данного объекта в ситуации актуальной коммуникации, имеют свою семантику. В частности, в русском языке жестов наличие угла между направлением глаз (взгляда) и направлением головы обычно свидетельствует о неискренности человека, о сокрытии им какой-то информации или нежелании сообщать её собеседнику, а тем самым и о недоверии к нему или недружелюбии. Ср. выражения *отводить глаза в сторону*, *опустить глаза*, *не смотреть <на собеседника>*, а также фразы (13) *Смотри мне в глаза и говори правду!*; (14) *Ты почему опустил глаза — говоришь неправду*? Во фразе (14),

которую произнесла учительница лицея в адрес ученика 6-ого класса, судя по интонации, явно выражена причинно-следственная связь между произнесением лжи и опусканием глаз.

Сходный смысловой анализ связи пространственных аспектов поведения человека и его отношения к собеседнику можно предложить также для жеста **смотреть исподлобья <на кого-л.>**. В самом деле, при его исполнении глаза жестикулирующего обращены на адресата, а голова — лицом чуть вниз и, возможно, вбок. Здесь тоже имеется небольшой угол между ориентациями глаз и головы, и, как и в разобранном выше случае, появление переносного значения у слова *исподлобья*. Ср. *исподлобья* = «из-под насупленных бровей <…> (также перен.: недоверчиво, недружелюбно)» [Ожегов 1983], «недоверчиво, недружелюбно; из-под насупленных, нахмуренных бровей (смотреть, глядеть)» [Кузнецов 1998]. Указанное переносное значение получает аналогичное объяснение: все перечисленные смыслы объединяют компоненты ‘неискренность’, ‘недружелюбие’. Общность смысловых компонентов наречий *искоса* и *исподлобья* позволяет им легко употребляться в одной синтагме, однако при этом каждое несёт свой смысловой акцент (при наличии общего смыслового инварианта, закреплённого за углом между телесными ориентациями): (15) *Смотрела на Клавдию* ***искоса, исподлобья****, как собака на строгого хозяина*; (16) *Еще робче сделала она первый глоток и, несмотря на сильный аппетит, приостановилась на минуту и глянула на меня* ***искоса, исподлобья****, желая поверней удостовериться, не намерен ли я тотчас же выкинуть над ней какую-нибудь скверную штуку. Так точно, с такими же приемами и почти с таким же выражением берут голодные, бездомные и запуганные собаки кусок пищи, брошенный рукой близко стоящего, незнакомого им человека*; (17) *Мало-помалу она приучилась на него смотреть, сначала* ***исподлобья, искоса****, и всё грустила, напевала свои песни вполголоса, так что, бывало, и мне становилось грустно, когда слушал её из соседней комнаты*.

По-видимому, не случайно В. И. Даль поясняет одно из значений слова *исподлобья* при помощи слова *искоса*: «непрямо, насупив брови и не поворачивая головы, искоса, насупясь; глядеть недоверчиво или со скрытным неудовольствием» [Даль 1998]. Хотя нормой общения русских людей в диалоге лицом к лицу является взгляд, направленный на собеседника, это — взгляд не прямо в глаза, поскольку жест **прямой взгляд в глаза** может интерпретироваться адресатом как вызов и потому плохо им восприниматься[5].

Подводя итог, мы можем сказать, что в том случае, когда один собеседник не смотрит на другого, как, например, в жесте **отвернуться** или **отвести глаза**, его невербальное поведение является показателем определённых психологических особенностей человека или черт его характера. В частности, поведение жестикулирующего может говорить о том, что он — очень стеснительный человек или что он не хочет быть назойливым, а может быть, просто не желает вступать в коммуникацию, поскольку ему неприятен либо сам собеседник, либо какие-то его качества.

---

[5] См. об этом [Kendon 1967]; [Von Cranach 1971]; [Ellsworth 1975]; [Emery 2000]; [Крейдлин 2002].

В таких жестах приветствия, как **протянуть руку для рукопожатия** или **протянуть руку для поцелуя**, определённый смысл приходится на компонент физической реализации «рука вытянута прямо». Любое отклонение от этой нормы исполнения жестов (например, протягивание руки ниже, чем прямо) свидетельствует о стремлении жестикулирующего управлять телесным поведением собеседника, в частности, о желании унизить его, подчеркнуть заметное различие в статусах и др. Такое пространственное поведение жестикулирующего лучше всего описывается посредством линейного представления ориентации, которое здесь выглядит следующим образом: рука жестикулирующего ориентирована кистью на адресата (не вниз и не вверх).

4. Заключение

При обсуждении двух возможных представлений пространственной ориентации (в виде конструкции линейной ориентации X V Y-ом Prep Z и конструкций угловой ориентации X V *под углом* Q Prep Z и X V Y-ом *под углом* Q Prep Z) и разных способов их реализаций мы особо остановились на соответствии значений телесного признака «ориентация» и значений психологического признака «отношение между собеседниками». Оба признака играют важнейшую роль в семиотической концептуализации тела и телесности, то есть в модели, или картине мире, отражающей представления обыкновенных носителей русского языка и русской культуры о человеческом теле, выраженные в знаках соответствующих семиотических кодов. Отношения между людьми могут быть как позитивными, так и негативными, как постоянными, так и временными, актуальными. Создание полного перечня таких отношений и их соответствий не только значениям признака «ориентация соматического объекта», но и значениям других физических и функциональных телесных признаков — это важная задача дальнейших исследований.

Литература

1. Апресян 2008 — *Апресян Ю. Д.* О проекте активного словаря (АС) русского языка // Компьютерная лингвистика и интеллектуальные технологии (по материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2008 г.)), Вып. 7 (14). М.: РГГУ, 2008. С. 23–31.
2. Даль 1998 — *Даль В. И.* Толковый словарь живого великорусского языка, в 4 т. М.: Терра, 1998.
3. Крейдлин 2002 — *Крейдлин Г. Е.* Невербальная семиотика. Язык тела и естественный язык. М.: Новое литературное обозрение, 2002.
4. Крейдлин 2012 — *Крейдлин Г. Е.* Тело человека и некоторые особенности синтаксического взаимодействия вербальных и невербальных знаковых кодов в диалоге // Современные прагмалингвистические и лингвокультурологические исследования. Иваново: Ивановский государственный университет, 2012. С. 144–152.

5. Крылова 2010 — *Крылова Т. В.* Лексикон «отвода глаз» (*отвернуться*, *отвести глаза*, *опустить глаза*, *потупиться*) // Логический анализ языка. Моно-, диа-, полилог в разных языках и культурах. М.: Индрик, 2010. — С. 184–195.
6. Кузнецов 1998 — Большой толковый словарь русского языка / Гл. ред. *С. А. Кузнецов*. СПб.: Норинт, 1998.
7. Морозова 2006 — *Морозова Е. Б.* Невербальный этикет в его соотношении с вербальным. Дисс. ... канд. филол. наук. М., 2006.
8. Ожегов 1983 — *Ожегов С. И.* Словарь русского языка, изд. 14-ое, стереотипное. М.: Русский язык, 1983.
9. Переверзева 2013 — *Переверзева С. И.* Семиотическая концептуализация тела в русском языке и русской культуре: признак «ориентация». Дисс. ... канд. филол. наук. М., 2013.
10. Подлеская, Рахилина 2000 — *Подлесская В. И.*, *Рахилина Е. В.* «Лицом к лицу» // Логический анализ языка: Язык пространства. М.: Языки русской культуры, 2000. С. 98–107.
11. Рахилина 2000 — *Рахилина Е. В.* Когнитивный анализ предметных имён. М.: Языки русской культуры, 2000.
12. Ellsworth 1975 — *Ellsworth P. S.* Direct gaze as a social stimulus: the example of aggression // Nonverbal communication of aggression. New York, 1975. P. 53–76.
13. Emery 2000 — *Emery N. J.* The eyes have it: The neuroethology, function and evolution of social gaze. Neuroscience and Biobehavioral Reviews, 24, 2000. P. 581–604.
14. Kendon 1967 — *Kendon A.* Some Functions of Gaze Direction in Social Interaction // Acta Psychologica, 26, 1967. P. 22–63.
15. Lakoff, Johnson 1980 — *Lakoff G., Johnson M.* Metaphors we live by. Chicago: University of Chicago Press, 1980.
16. Von Cranach 1971 — *Von Cranach M.* The role of orienting behavior in human interaction // A. H. Esser (ed.) Behavior and environment: the use of space by animals and men. NY: Plenum Press, 1971. P. 217–237.

References

1. *Apresjan Ju. D.* (2008), On the project of the active dictionary of Russian [O projekte aktivnogo slovarja (AS) russkogo jazyka], Computational Linguistics and intellectual technologies [Kompjuternaja lingvistika i intellektual'nye tehnologii (po materialam ezhegodnoj mezhdunarodnoj konferentsii "Dialog")], Vol. 7 (14), pp. 23–31.
2. *Dal' V.I.* (1998), The explanatory dictionary of the living great Russian Language [Tolkovyj slovar' zhivogo velikorusskogo jazyka], Terra, Moscow.
3. *Kreydlin G. E.* (2002), Nonverbal semiotic: body language and natural language [Neverbal'naja semiotika: jazyk tela i estestvennij jazyk], Novoe literaturnoe obozrenie, Moscow.
4. *Kreydlin G. E.* (2012) The human body and some peculiarities of syntactic interaction between verbal and nonverbal sign codes in the dialog [Telo

cheloveka i nekotoryje osobennosti sintaksicheskogo vzaimodejstvija verbal'nyh i neverbal'nyh znakovyh kodov v dialoge], Sovremennye pragmalingvisticheskie i lingvokul'turologicheskie issledovanija, Ivanovo State University, Ivanovo, pp. 144–152.

5. *Krylova T. V.* (2010), The lexicon of "eye aversion" (*to turn away, to avert one's eyes*, *to cast one's eyes down*) [Leksikon "otvoda glaz" (*otvernut'sja*, *otvesti glaza*, *opustit' glaza*, *potupit'sja*)], Logical analysis of the language. Mono-, dia-, polylog in different languages and cultures [Logicheskij analiz jazyka. Mono-, dia-, polilog v raznyh jazykah i kulturah], Indrik, Moscow, pp. 184–195.
6. *Kuznetsov S. A.* (1998), The big explanatory dictionary of the Russian language [Bol'zhoj tolkovyj slovar' russkogo jazyka], Norint, Saint-Petersburg.
7. *Morozova E. B.* (2006), Nonverbal etiquette and its relations with the verbal etiquette [Neverbal'nyj ètiket v ego sootnoshenii s verbal'nym], Moscow.
8. *Ozhegov S. I.* (1983), The dictionary of the Russian language [Slovar' russkogo jazyka], 14-th stereotype edition, Russkij jazyk, Moscow.
9. *Pereverzeva S. I.* (2013), The semiotic conceptualization of the body in the Russian language and culture: the feature of orientation [Semioticheskaja kontseptualizatsija tela v russkom jazyke i russkoj culture: priznak "orientatsija"], Moscow.
10. *Podlesskaja V. I., Rahilina E. V.* (2000), "Face to face" ["Litsom k litsu"], Logical analysis of the language. The language of space [Logicheskij analiz jazyka. Jazyk prostranstva]. Jazyki russkoj kul'tury, Moscow, pp. 98–107.
11. *Rahilina E. V.* (2000) The cognitive analysis of physical names], Jazyki russkoj kul'tury, Moscow.
12. *Ellsworth P. S.* (1975), Direct gaze as a social stimulus: the example of aggression, Nonverbal communication of aggression, New York, pp. 53–76.
13. *Emery N. J.* (2000), The eyes have it: The neuroethology, function and evolution of social gaze. Neuroscience and Biobehavioral Reviews, 24, pp. 581–604.
14. *Kendon A.* (1967), Some Functions of Gaze Direction in Social Interaction, Acta Psychologica, 26, pp. 22–63.
15. *Lakoff G., Johnson M.* (1980), Metaphors we live by, Chicago, University of Chicago Press.
16. *Von Cranach M.* (1971), The role of orienting behavior in human interaction, Behavior and environment: the use of space by animals and men, New York, Plenum Press, pp. 217–237.

A DATABASE OF RUSSIAN VERBAL FORMS AND THEIR FRENCH TRANSLATION EQUIVALENTS[1]

Kruzhkov M. G. (magnit75@yandex.ru)
IPI RAS, Moscow, Russia

Buntman N. V. (nabunt@hotmail.com)
MSU, Moscow, Russia

Loshchilova E. Ju. (lena0911@mail.ru)
IPI RAS, Moscow, Russia

Sitchinava D. V. (mitrius@gmail.com)
IRL RAS, Moscow, Russia

Zalisniak A. A. (anna.zalizniak@gmail.com)
IL RAS, IPI RAS, Moscow, Russia

Zatsman I. M. (izatsman@yandex.ru)
IPI RAS, Moscow, Russia

The paper presents the results of a project[2] aimed at the development of methodology and information technology for the creation of a corpus-based linguistic database of verbal forms with their translation equivalents (with bilingual grammatical search functions). Within the scope of the project the following results have been achieved:

1. *Methodology and information technology* for the creation of linguistic databases based on bilingual parallel corpora have been developed (including corpora with multiple translation variants).

2. The *polyvariant parallel subcorpus* which includes Russian literary works with French translations has been created within the Russian National Corpus (RNC). Some of the parallel texts in the subcorpus include multiple translation variants.

3. On the basis of the polyvariant Russian-French corpus a *database of Russian verbal lexico-grammatical forms* (LGFs) and their French translation equivalents has been created. Equipped with bilingual grammatical search functions, the database is a unique resource that can be used for investigating a wide range of various cross-linguistic problems.

4. A number of concepts in the areas of Russian verbal categories and Russian-French contrastive grammar have been refined.

Key words: polyvariant parallel corpus, grammatical semantics, contrastive grammar, database of verbal forms, corpus linguistics, bilingual grammatical search, Russian, French

---

[1] The work was carried out at the Institute of Informatics Problems of the RAS.

[2] The project "Information technology for creation of corpus-based linguistic database of verbal forms with their functional equivalents" was supported by "Dinastija" foundation, grant NG13-036.

1. Introduction

The emergence of electronic corpora has marked the beginning of a new era in contrastive linguistics. Stig Johansson's work with English-Norwegian parallel corpus in 1990s was ground-breaking in this field [1]. Combining the methodological advantages of computer corpus linguistics with the possibility of contrasting parallel texts in two or more languages allowed to compare the actual use of the languages involved at all levels of descriptions—with greater accuracy and detail than had been possible before. During the past two decades considerable advances have been made in this field, both in development of analytical methods and in creation of unique lexicographic descriptions. Further avenues of research in contrastive grammatical studies are discussed in the following works: [1, 2, 3, 4, 5, 6, 7].

The technology of compilation of the Russian-French polyvariant parallel subcorpus of the Russian National Corpus (RNC) is presented in [8]. The subcorpus contains Russian literary works aligned with French translations with current total volume of 2 million words. Some of the parallel texts are offered in *the polyvariant format*, i.e. a single text in Russian is aligned with more than one translation of the same text into French. Total volume of the polyvariant texts is currently 700 thousand words—more than a third of the total volume of the Russian-French subcorpus of the RNC.

Parallel subcorpora were first introduced into the RNC in 2005 [2, 4, 9, 10]. Currently the RNC includes 8 bilingual parallel subcorpora with Russian as source or target language: English, German, French, Spanish, Italian, Polish, Ukrainian and Byelorussian ones. There is also one multilingual parallel subcorpus. The Russian-French parallel subcorpus was added to the RNC in December 2012. The technology used for the compilation of this subcorpus (presented in [8]) hereafter is referred to as Parallel Corpus technology or PC-technology.

In 2013 PC-technology was extended with the introduction of new operations designed for creation of *the Database of Russian Verbal Forms and their French Translation Equivalents* (hereafter—the DB)[3] and for the compilation of the polyvariant Russian-French subcorpus (hereafter—the subcorpus). The extended technology allows to simultaneously compile the subcorpus and to fill the DB. In addition the extended technology implements bilingual grammatical search functions for the verbal forms and their translations in the DB. For instance, in the DB a user can browse all Russian verbal forms in Russian present tense which are translated into French by *passé composé*. The new extended technology hereafter is referred to as Database Parallel Corpus technology or DBPC-technology.

The goal of this paper is to describe the purpose of the DB, which is based on the parallel texts from the subcorpus, and to illustrate the bilingual grammatical search function implemented in the DB.

---

[3] The list of the database contributors is as follows: N. Buntman, B. Loktev, V. Nuriev, O. Petrushkina, N. Popkova, E. Roganova, E. Spiridonova, V. Stepanov, A. Shchurova.

2. Purpose and methodology

The methodology of the database construction is primarily defined by its purpose. The DB was created as a tool that should allow to describe Russian grammatical semantics "as mirrored in French" and to clarify certain concepts in Russian-French contrastive grammar. In the methodology development we relied on works by V. G. Gak, I. N. Kouznetsova, M. Guiraud-Weber [11, 12, 13, 14] and other authors. But these works were created in the pre-corpus era; today when Russian-French corpora are regularly compiled and updated, we can rely on parallel texts analysis to update descriptions of Russian-French contrastive grammar.

While developing the DB we kept in mind that the main object of analysis for linguistic experts working with the DB was the correspondences between Russian and French verbal categories in parallel texts. In order to properly describe the analyzed correspondences, a number of terms have been defined capturing the essence of the developed methodology [15, 16, 17]

Among the key notions are ***lexico-grammatical form*** (**LGF**) and ***basic LGF type***. Basic LGF type is understood as a certain *combination of grammatical features* along with certain elements in the sentence structure which define a certain "construction"[4]; consider, for example, basic LGF type "PastPF + *если бы*" (= past tense, perfective aspect + *если бы*); cf. also 3nd and 5th column on Fig.1. Accordingly, LGF is understood as a *combination of elements of a sentence* which realizes a given set of features, for example (elements of the LGF are marked in bold): ***если бы*** *он* ***пришел*** *вовремя*; cf. also 2nd and 4th column on Fig. 1.

For Russian 15 basic LGF types were specified (see Table 1)[5]. This is the so-called Source Set that restricts the initial search of Russian LGFs in the DB on the first stage of DBPC-technology. The scope of French basic LGF types is not limited: it is continually expanded as the new types of translation variants are identified in the DB. At present, the experimental version of the DB includes 25 French basic LGF types (see Table 2).

Table 1. The Source Set: basic verbal LGF types (Russian)

1.	Present	Pres-IPF
2.	Past Imperfective	Past-IPF
3.	Past Perfective	Past-PF
4.	Simple Future	Fut-PF
5.	Compound Future	Fut-IPF
6.	Imperative Perfective	Imperat-PF
7.	Imperative Imperfective	Imperat-IPF
8.	Form with *бы* PF	Past-PF+*бы*

[4] As this term is understood in the Construction Grammar [15–17].

[5] The actual DB includes only LGFs with finite verbal forms (i.e. impersonal verbs, participles, periphrases with the verb *быть* are not included). In the future the range of examined types of verbal forms will be expanded.

9.	Form with *бы* IPF	Past-IPF+*бы*
10.	Form with *если бы* PF	Past-PF+*если бы*
11.	Form with *если бы* IPF	Past-IPF+*если бы*
12.	Form with *чтобы* PF	Past-PF+*чтобы*
13.	Form with *чтобы* IPF	Past-IPF+*чтобы*
14.	Form with *было* PF	Past-PF+*было*
15.	Form with *было* IPF	Past-IPF+*было*

Table 2. The Target Set: basic LGF types (French)

1.	présent	Pr
2.	passé composé	PasCom
3.	passé simple	PasSim
4.	imparfait	Imparf
5.	plus-que-parfait	PqParf
6.	passé antérieur	PasAnt
7.	passé immédiat	PasIm
8.	futur simple	Fut
9.	futur antérieur	FutAnt
10.	futur immédiat	FutIm
11.	impératif	Imperat
12.	subjonctif présent	SubjPres
13.	subjonctif passé	SubjPas
14.	subjonctif imparfait	SubjImparf
15.	subjonctif plus-que-parfait	SubjPqParf
16.	conditionnel présent	CondPr
17.	conditionnel passé	CondPas
18.	participe présent	PartPr
19.	participe passé	PartPas
20.	participe passe compose	PartPasComp
21.	gérondif	en PartPr
22.	infinitif	Inf
23.	préposition+infinitif	Prep+Inf
24.	préposition+infinitif passé	Prep+InfPas
25.	substantif	Subst

Apart from the features that define the basic LGF types, lists of additional features has been compiled for each of the two languages. Additional features allow to make a further specification of the type of construction. They define either the composition of the verbal group (e.g. presence of a subordinate infinitive, a modality marker, a negation marker), or the type of the sentence in which the LGF is used (e.g. subordinate clause, interrogative sentence, direct speech), see Tables 3 and 4. Each additional feature can apply to one or more basic LGF types. On Figures 1, 2, 5, 6

additional features are specified in square brackets after the basic LGF type. *LGF type* is defined as combination of a basic LGF type and a set of relevant additional features.

Table 3. Additional features for basic LGF types (Russian)

Subordinate infinitive PF	[SubInf-PF]
Subordinate infinitive IPF	[SubInf-IPF]
Modality marker	[ModDet]
Negation	[Neg]
Interrogative sentence	[Interrog]
Exclamatory sentence	[Exclam]
Verb introducing direct speech	[VerbDirSp]
Verb inside direct speech	[DialRepl]
Verb in complement clause	[SubCompl]
Verb in attributive clause	[SubAttr]
Verb in subordinate clause	[Sub]

Table 4. Additional features for basic LGF types (French)

Subordinate infinitive	[SubInf]
Subordinate past infinitive	[SubInfPas]
Subordinating predicate added	[+SuperPred]
Modality marker	[ModDet]
Negation	[Neg]
Interrogative sentence	[Interrog]
Exclamatory sentence	[Exclam]
Verb introducing direct speech	[VerbDirSp]
Verb inside direct speech	[DialRepl]
Verb in complement clause	[SubCompl]
Verb in attributive clause	[SubAttr]
Verb in subordinate conditional clause	[SubCond]
Verb in subordinate clause	[Sub]
Accusativus cum infinitivo	[Acc.c.Inf]

The process of establishing correspondences between Russian and French LGFs in aligned parallel texts is carried out as follows. First, an expert marks in the Russian text a fragment corresponding to one of the 15 *basic LGF types* from the Source Set, i.e. one Russian *LGF*. Then, the expert looks for its "functionally equivalent fragment" (FEF)[6] in the aligned translated fragment, marks it and matches it to an appropriate *basic LGF type* in the Target Set. If the needed unit is not listed in the Table 2, the Target Set can be expanded. If a FEF cannot be located in the translation for a certain

[6] The term "functionally equivalent fragment" was introduced in [2, 9].

Russian LGF, the ME is marked as "Nondetermined" in the DB and is not taken into account when processing the data[7].

A pair of fragments obtained in this way is referred to as a *monoequivalence* (ME, see definition below), see Fig. 1. Extraction of LGFs and FEFs and matching them to appropriate LGF types and to each other is the initial task that is supported by the DB. Before we can pass on to further tasks, we have to define a few new terms in the area of contrastive analysis: *monoequivalence* (ME), *type of ME*, *polyequivalence* (PE), *type of PE* and *hyperequivalence* (HE).

ME#	Russian LGF	Russian LGF type	French LGF	French LGF type
4711	потом [...] плотно **запер** все двери	Past-PF [ModDet]	après **avoir** bien **fermé** toutes les portes.	InfPas [Prep+InfPas] [Sub]

Fig. 1. A sample ME from the DB

Monoequivalence is a pair *<Rn(i); Fm(j)>,* where *Rn(i)* is a specific LGF of Russian basic LGF type *Rn* (see Table 1) in the original text, and *Fm(j)* is a specific LGF of French basic LGF type *Fm* (see Table 2) in one of the translations. All LGFs in the DB have unique identifiers, so in this case specific LGFs are uniquely identified by indexes *i* and *j*.

Type of ME *<Rn(i); Fm(j)>* is the pair of the corresponding basic LGF types *<Rn; Fm>,* e.g. for the ME represented on the Fig.1 it is: <Past-PF; InfPas>.

Polyequivalence *<Rn(i)*; *{Fm(j), Fk(r), …}>* is a combination of monoequivalences *<Rn(i)*; *Fm(j)>*, *<Rn(i)*; *Fk(r)>* etc. with identical Russian LGF in the first position. A PE reflects different variants of translation of the same original LGF (see Fig. 2).

Type of PE *<Rn(i)*; *{Fm(j), Fk(r), …}>* is the pair *<Rn*; *{Fm, Fk, …}>* e.g. <Present; {Présent, Présent}> (see 2nd and 5th columns of Fig. 2).

Hyperequivalence is a pair *<Rn; {F}>*, which represent aggregation of all possible *types of ME* in the DB with the same value at the first position. It comprises of one Russian basic LGF type *Rn* and a multitude of French basic LGF types *{F}* that enter into MEs with Russian LGFs of basic LGF type *Rn*.

Based on the terms defined above we enumerate the tasks that the developed DB is meant to accomplish:

- building of MEs, PEs and HEs;
- bilingual grammatical search of MEs and PEs;
- calculating frequencies for each type of ME and PE in the DB.

---

[7] Here we refer to such cases when the lexical items used to translate the semantic content enclosed in the original LGF are substantially different from the original, to the extent that it is impossible to establish a correspondence between the LGFs using the existing apparatus. E.g.: *ты [...] так теребишь за носы, **что еле держатся**—tu tirais tellement sur leur nez [...] que **tu as failli le leur arracher***.

Russian LGF	Russian LGF type	LGFs in French translations		
		ME#	French LGFs	French LGF Types
Я иногда в театр **хожу**	Pres-IPF [ModDet] [DialRepl]	596	**Il m'arrive d'aller** au théâtre,	Pr [SubInf] [+SuperPred] [DialRepl]
		5927	Non, je **vais** parfois au théâtre, et en visite.	Pr [ModDet] [DialRepl]

Fig. 2. A sample PE combined from two MEs[8]

In order to facilitate accomplishment of these tasks we have developed a web interface that allows users (linguistic experts) to interact with the DB using the common web browsers (Internet Explorer, Mozilla Firefox, Google Chrome).

The DB functions can be divided into two major groups: the first group supports building and editing of MEs (see Fig. 3), and the second group supports viewing and searching of MEs and PEs (for PE search functions see Fig. 4).

Building and editing functions allow filtering of the aligned fragments of Russian and French texts by book title, translator and basic LGF types that the user looks for in the original text. The users then browse the selected pairs of aligned fragments in order to find LGFs and to build MEs.

By the beginning of 2014, 10527 MEs have been built, and 4128 PEs have been automatically generated by matching MEs with the same Russian LGF at the first position.

3. Bilingual grammatical search functions

The PE search page allows users to view collections of PEs (Fig. 4) that are generated according to the specified search queries. Users can filter all the PEs in the DB using such search features as original book title, Russian and French LGF types, specific lexemes or text fragments in the original text or in the translation, etc. The search features can be specified separately or in any combinations. After the query is executed, the user can see the number of selected PEs and browse the found PEs.

[8] The French LGFs in this PE belong to the same basic LGF type but they have different additional features.

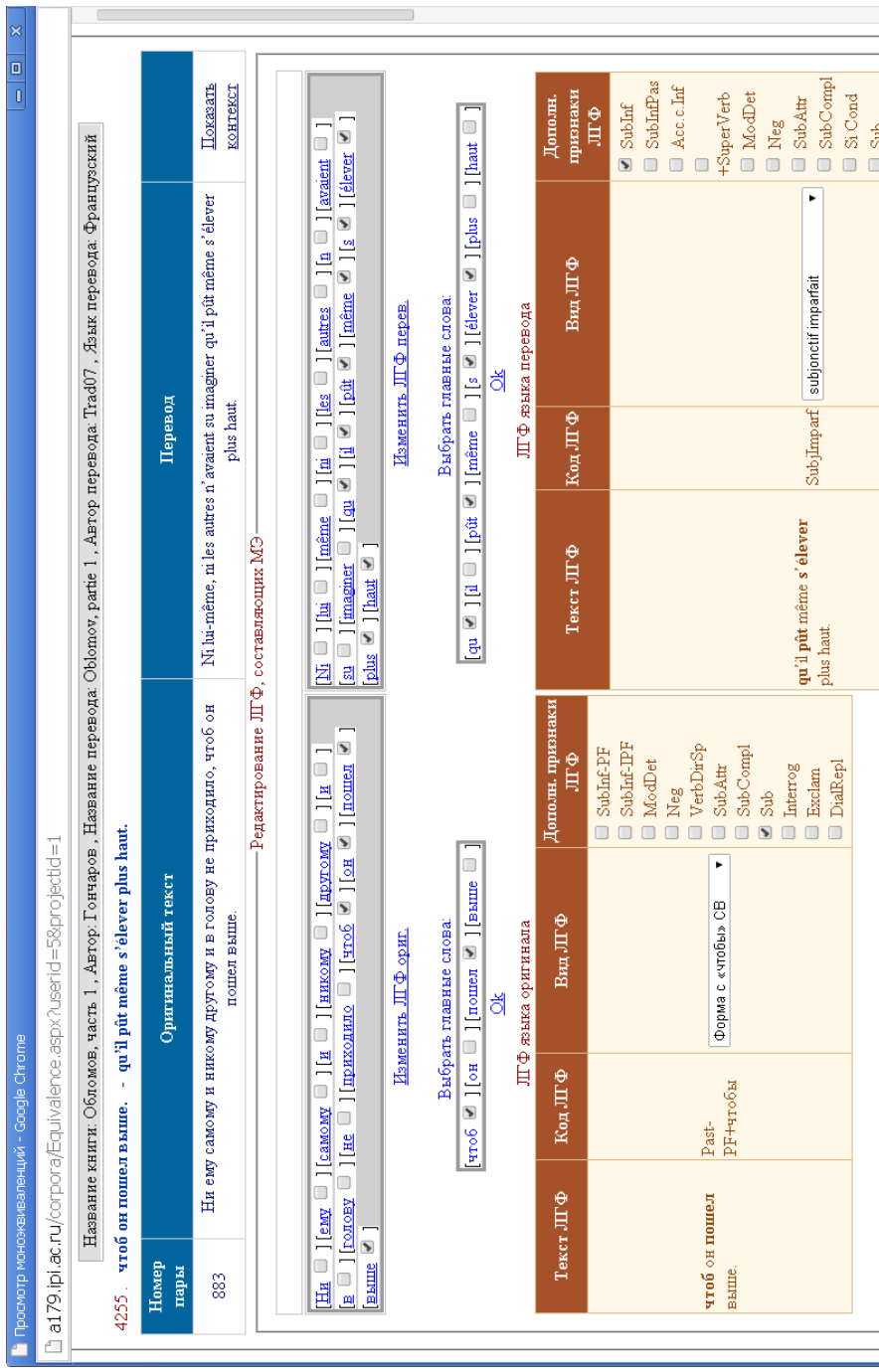


Fig. 3. Web interface for editing MEs

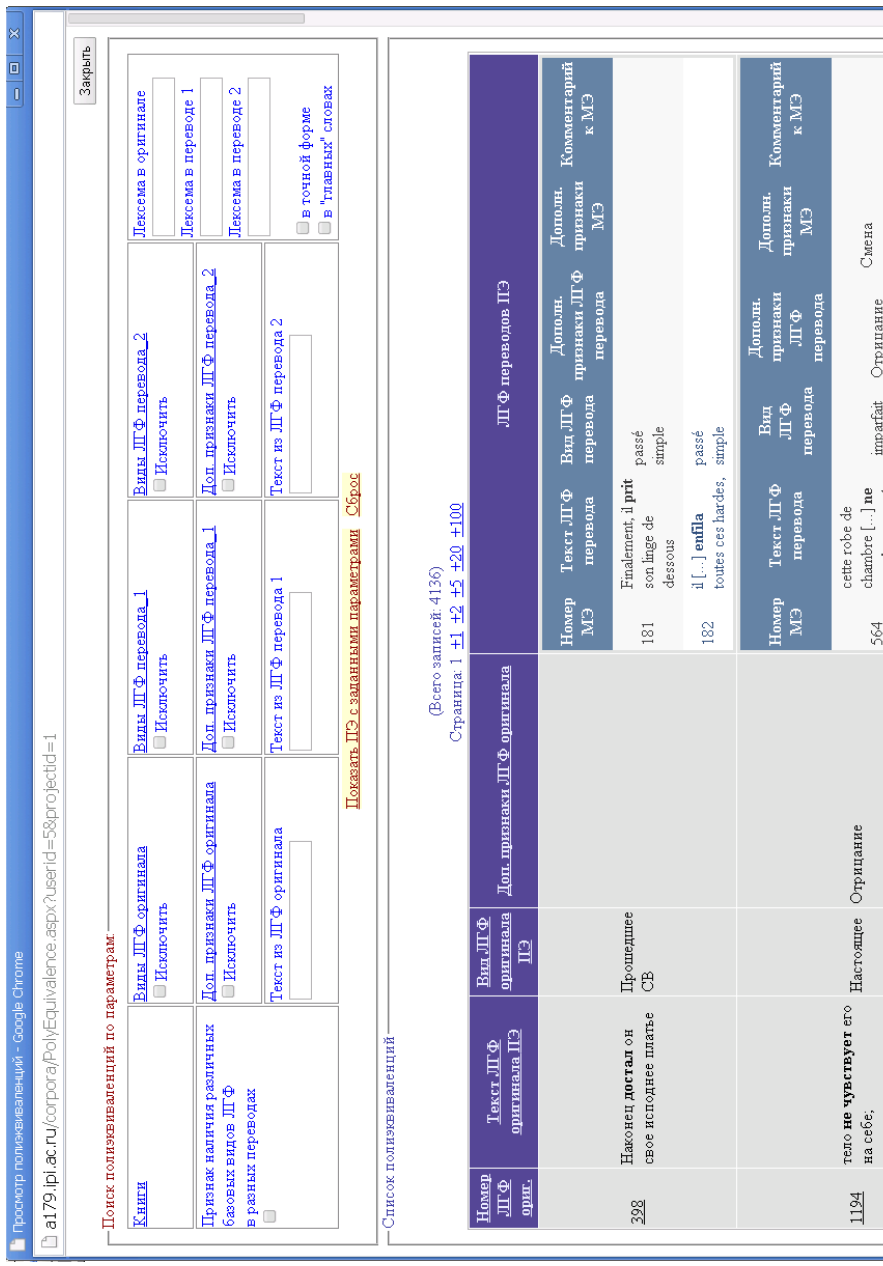


Fig. 4. Web interface for viewing and searching PEs

The bilingual grammatical search which can be applied to one or more translations (a polyvariant bilingual query) is a fundamentally new research tool. For instance, we can specify a Russian basic LGF type (*Past-PF*) and two different French

basic LGF types for two translation variants (*CondPr* and *PasSim*). Such a query will result in two PEs being found (see Fig. 5).

		ME#	Translation LGF	Translation LGF type
он **решил оставить** [...] липовые и дубовые деревья	Past-PF [SubInf]	2931	alors qu'**il garderait** les [...] tilleuls et chênes,	CondPr
		8011	Il **décida de laisser** tels quels les [...] tilleuls et les chênes,	PasSim [SubInf]
он **решил** [...] яблони и груши **уничтожить**	Past-PF [SubInf]	ME#	Translation LGF	Translation LGF type
		2932	il **se débarrasserait** des pommiers et des poiriers	CondPr
		8013	Il **décida** [...] **de supprimer** les pommiers	PasSim [SubInf]

Fig. 5. 2 PEs found in the DB by specifying the basic LGF types and two translation variants

Apart from the basic LGF type, the user can specify additional features of Russian and French LGFs (see Tables 3 and 4). For instance, we can specify Russian LGF type *Pres [SubInf-PF]* and French LGF Type *CondPr [SubInf]* in at least one of the translation variants. Such a query will result in three PEs being found (see Fig 6).

4. Conclusion

The created DB allowed us to clarify some concepts in Russian-French contrastive grammar. In particular, the list of correspondences described in [11, 12] and summarized in [13] has been:

- inverted (in works by Gak and Kouznetsova language material was examined from the viewpoint of translating French texts into Russian because their goal was interpretation of meanings and functions of French forms);
- expanded, i.e. new types of translation correspondences have been established.

Of particular interest are the results of the frequency analysis of translational correspondences. For example, correlation between oppositions "perfective *vs.* imperfective aspect" in Russian and "passé composé/passé simple *vs.* imparfait" in French can be refined based on quantitative indicators. Russian basic LGF type Past-PF only in 49,4% of cases corresponds to French basic LGF type Imparf and in 21% of cases to PasCom/PasSim. These figures highlight the width of the semantic range of the Russian imperfective aspect.

		ME#	Translation LGF	Translation LGF type
Не может постараться для барина!	Pres [SubInf-PF] [Neg]	661	**Tu pourrais** tout de même **faire un effort** pour ton maître!	CondPr [SubInf] [Exclam]
		5897	Il **ne peut** même **pas faire** un petit effort pour son maître!	Présent [SubInf] [Exclam]
теперь **можете отдать**	Pres [SubInf-PF]	945	maintenant vous **pouvez** me **rembourser**.	Présent [SubInf]
		7584	alors vous **pourriez** peut-être **me rembourser**?	CondPr [SubInf] [ModDet] [Interrog]
Разве я **могу** все это [...] **перенести**?	Pres [SubInf-PF] [Interrog]	8939	Est-ce que je **puis** [...] le **supporter**?	Présent [SubInf] [Interrog]
		8940	Je **pourrais** [...] **supporter** tout ça?	CondPr [SubInf] [Interrog]

Fig. 6. 3 PEs found in the DB by specifying the basic LGF types with additional features

Furthermore, the DB based on the polyvariant subcorpus allows to clarify the semantics of Russian verbal forms: French translation variants with more detailed network of grammatical oppositions in the domain of tense/mood features make it possible to detect specific semantic components in Russian LGFs.

The DB creation has confirmed the efficiency of "construction" as a tool of linguistic analysis: the contrastive approach based on the term "LGF" allows to bring to light various relationships between actional, temporal, aspectual and modal components in meanings of Russian verbal forms.

Finally, the developed DBPC-technology is readily adoptable to be used in other cross-linguistic projects based on parallel aligned texts (i.e. projects dedicated to investigation of other categories of LGFs, not necessary verbal ones). The customization can be accomplished without significant changes to the structure of the DB. To customize the DBPC-technology one should basically supply the list of languages used and the lists of basic LGF types and additional features for these languages according to goals and objectives of a specific project.

Currently the DBPC-technology is being adopted for investigation of Russian language-specific units.

References

1. *Aijmer K., Altenberg B.* (2013), Advances in Corpus-based Contrastive Linguistics. Studies in honour of Stig Johansson, John Benjamins, Amsterdam.
2. *Dobrovol'skij D. O., Kretov A. A., Sharov S. A.* (2005), Parallel text corpus [Korpus parallel'nyh tekstov], Scientific and technical information. Ser.2. Information processes and systems [Nauchnaja i tehnicheskaja informacija. Ser. 2. Informacionnye processy i sistemy], Vol. 6, pp. 16–27.
3. *Kiseleva K. L., Plungjan V. A., Rahlina E. V., Tatevosov S. G.* (Eds.) (2009), Corpus-based research on Russian grammar [Korpusnye issledovanija po russkoj grammatike], Probel-2000, Moscow.
4. *Dobrovol'skij D. O.* (2009), Parallel text corpus in studying of culture-specific words [Korpus parallel'nyh tekstov v issledovanii kul'turno-specifichnoj leksiki], Russian National Corpus: 2006–2008. New results and prospects [Nacional'nyj korpus russkogo jazyka: 2006–2008. Novye rezul'taty i perspektivy], Nestor-Istorija, St. Petersburg, pp. 383–401.
5. *Sichinava D. V.* (2011), Complex study of monolingual and parallel text corpora in grammatical studies [Kompleksnoe issledovanie odnojazychnogo i parallel'nogo korpusov v grammaticheskih issledovanijah], Proceedings of the international conference "Corpus Linguistics—2011" [Trudy mezhdunarodnoj konferencii «Korpusnaja lingvistika—2011»], St. Petersburg, pp. 316–322.
6. *Sichinava D. V., Shvedova M. A.* (2010), Parallel corpora within Russian National Corpus: technologies and the addressed issues [Parallel'nye korpusa v sostave Nacional'nogo korpusa russkogo jazyka: tehnologii i reshaemye zadachi], Computational linguistics: areas of study and research [Komp'juternaja lingvistika: nauchnoe napravlenie i uchebnaja disciplina], Gomel', GGU imeni F. Skoriny, pp. 30–34.
7. *Sichinava D. V., Arhangel'skij T. A.* (2012), Belarusian-Russian and Russian-Belarusian parallel corpora: Russian National Corpus collaboration [Parallel'nye belorussko-russkij i russko-belorusskij korpusa: sovmestnyj proekt Nacional'nogo korpusa russkogo jazyka], TEL-2012 workshop proceedings [Trudy shkoly-seminara TEL-2012], Kazanskij (Privolzhskij) federal'nyj universitet, Kazan'.
8. *Zalizniak Anna A., Sitchinava D. V., Loiseau S., Kruzhkov M., Zatsman I. M.* (2013), Database of Equivalent Verbal Forms in a Russian-French Multivariant Parallel Corpus, 2013 International Conference on Artificial Intelligence (ICAI'13) Vol. I, Las Vegas, pp. 101–107.
9. *Dobrovol'skij D. O., Kretov A. A., Sharov S. A.* (2005), Parallel text corpus: architecture and possible application [Korpus parallel'nyh tekstov: arhitektura i vozmozhnosti ispol'zovanija], Russian National Corpus: 2003–2005 [Nacional'nyj korpus russkogo jazyka: 2003–2005], Indrik, Moscow, pp. 263–296.
10. *Andreeva E. G., Kasevich V. B.* (2005), Grammar and Lexis (evidence from English-Russian parallel text corpus) [Grammatika i leksika (na materiale anglo-russkogo korpusa parallel'nyh tekstov)], Russian National Corpus: 2003–2005 [Nacional'nyj korpus russkogo jazyka: 2003–2005], Indrik, Moscow, pp. 297–307.
11. *Gak V. G.* (2006), The Russian language in comparison with the French language [Russkij jazyk v sopostavlenii s francuzskim], URSS, Moscow.

12. *Gak V. G.* (2009), Comparative typology of the French language and the Russian language [Sravnitel'naja tipologija francuzskogo i russkogo jazykov], URSS, Moscow.
13. *Kouznetsova I. N.* (2009), Contrastive Grammar of French and Russian [Grammaire contrastive du francais et du russe], Nestor Academic Publishers, Moscow.
14. *Guiraud-Weber M.* (2011), Essays on Russian and contrastive syntax [Essais de syntaxe russe et contrastive], Université de Provence, Aix-en-Provence.
15. *Rahilina E. V.* (red.) (2010), Construction linguistics [Lingvistika konstrukcij], Azbukovnik, Moscow.
16. *Goldberg A.* (1995), Constructions: A Construction Grammar Approach to Argument structure, Univ. of Chicago Press, Chicago.
17. *Goldberg A.* (2006), Constructions at Work. The nature of generalization in grammar, Oxford Univ. Press, Oxford.

CONDITIONAL RANDOM FIELD IN SEGMENTATION AND NOUN PHRASE INCLINATION TASKS FOR RUSSIAN

Kudinov M. S. (m.kudinov@samsung.com)

Samsung R&D Institute Russia, Moscow, Russia;
Dorodnitsyn Computing Center RAS, Moscow, Russia

Romanenko A. A. (a.romanenko@samsung.com)

Samsung R&D Institute Russia, Moscow, Russia;
Moscow Institute of Physics and Technology, Moscow, Russia

Piontkovskaja I. I. (p.irina@samsung.com)

Samsung R&D Institute Russia, Moscow, Russia

We propose solutions of several NLP problems for Russian making use of the conditional random fields (CRF) framework, including: shallow parsing (chunking), temporal expressions extraction and noun phrase inflection. Each of the three problems are important in speech generation, data mining and spoken dialogs systems design. The purpose of shallow parsing is to extract from the text syntactically related word forms (e.g. noun phrases) without full parsing. It may be useful in data mining applications. Temporal expressions extraction is important for natural language understanding modules of spoken dialog systems. Usually rule-based methods are used to address this problem. Noun phrase inflection is needed for speech generation modules. The main problem is to detect word forms for inflection. For all three problems statistical approach was taken. We use simple version of CRF named linear-chain CRF. In shallow parsing and time expressions extraction state-of-the-art results were achieved. In noun phrase inflection, the level of F_1-measure exceeded 95.

Key words: NLP, conditional random field, CRF, chunking, shalow parsing, temporal expressions extraction, noun phrase inflection

1. Introduction

It has been shown that a number of popular NLP tasks can be considered as the sequence labeling problem with certain output vocabulary. The typical ones are POS tagging, shallow parsing (chunking), temporal expression extraction and co-reference resolution. In bioinformatics and natural language processing the sequence labeling problem can be stated as finding the optimal mapping between an input sequence in an alphabet D onto a an output sequence in an alphabet L. During the past decades a number of both rule-based [1] and statistical [14] methods for solving this problem have been proposed. There is much evidence (e.g. [13]) of higher

performance demonstrated by linear-chain conditional random fields (L-CRF) as a statistical tool for machine learning algorithms. L-CRF is a discriminative model and in this aspect it resembles the popular Maximum entropy Markov model (MEMM). However, it was demonstrated ([3,7]) that MEMM has a considerable flaw named *label bias*. The problem is that the learning algorithm of MEMM causes the model to be more likely to choose hidden sates with lower entropy of transition probability distribution. For instance in POS tagging task MEMM will tend to choose those tags which prefer fewer types of followers. L-CRF was successfully applied for POS tagging in [7]. It was also successfully applied for shallow parsing in 13 and for co-reference resolution in [5].

Application of L-CRF to Russian is observed in [2]. In this paper POS tagging, co-reference resolution and sentiment analysis tasks have been considered.

In the present paper we propose a solution of three NLP tasks applied to Russian: temporal expression extraction, shallow parsing and inflection of noun phrases. We show that all these tasks may be reduced to sequence labeling problem and solved by means of L-CRF model. The remainder of the paper is organized as follows. In the next paragraph we give a short mathematical description of the L-CRF model; the next paragraphs are dedicated to the problems enumerated above. Each part includes description of the task, description of the datasets and the experimental results.

2. Linear-Chain CRF

CRF is a discriminative probabilistic graphical model defined induced by a non-oriented graph. The vertices of the graph correspond to random variables and edges correspond to probabilistic relations between them. In fact CRF is a graphical representation of a joint distribution $(Y_1, \ldots, Y_s)$ conditioned on observed data $(t_1, \ldots, t_s)$, where ti is a vector of observed features. Let Y_1^s and t_1^s be label and observation vector sequences from 1 to s correspondingly. Then we write the distribution of hidden label sequence conditioned on the observed feature vector sequence: $P(Y_1^s \mid t_1^s)$.

Consider the model in detail. The corresponding graph is given in Fig.1.

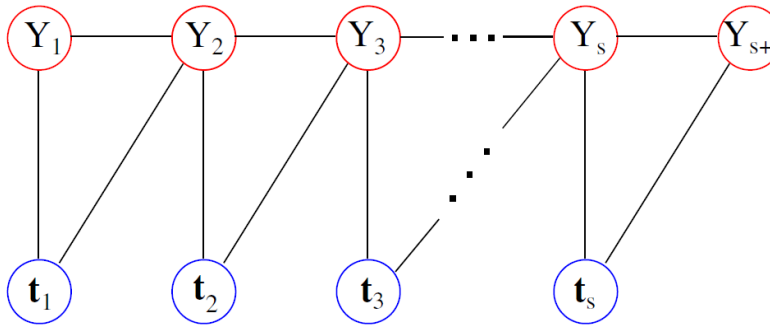


Fig. 1. The grahical model for a sentence of length *s*

Here $Y_i \in \{B, I, O\}, t_i \in \mathrm{T}$ are hidden label and observed feature vector on the i^{th} position, T stands for the set of allowable features.

In accordance with the model the sequence labelling task is formulated as finding the sequence $(\mathrm{Y}1, \ldots, \mathrm{Y}s)$, minimizing the distribution $P(Y_1^s \mid t_1^s)$. According to Hammersley-Clifford teorem such distribution may be factorized into function of arguments corresponding to the vertices of the graph specifying the CRF:

$$P(Y_1^s \mid t_1^s) = \frac{\prod_{i=1}^{s} \psi(Y_i, Y_{i+1}, t_i)}{Z} \tag{1}$$

where $Z = \sum_{Y_1^s} \prod_{i=1}^{s} \psi(Y_i, Y_{i+1}, t_i)$ is a normalizing factor or *partition function*.

$\psi(Y_i, Y_{i+1}, t_i)$ is a potential function of the graph clique calculated as:

$$\psi(Y_i, Y_{i+1}, t_i) = \exp \sum_{j=1}^{K} w_j f_j(Y_i, Y_{i+1}, t_i) \tag{2}$$

Here $f_j(Y_i, Y_{i+1}, t_i)$ is a j^{th} feature of the clique (Y_i, Y_{i+1}, t_i); K is a number of the features, $w_1 \ldots w_k$ are *feature weights* (model parameters). To find the model maximizing $P(Y_1^s \mid t_1^s)$ we maximize the sum:

$$\arg\max_{Y_1^s} \sum_{i=1}^{s} \sum_{j=1}^{K} w_j f_j(Y_i, Y_{i+1}, t_i) \tag{3}$$

To maximize (3) *Viterbi algorithm* can be used. The learning procedure is discussed in detail in [6].

3. Temporal Expressions Extraction

3.1. Task Definition

The task of temporal expressions extraction is a kind of named entity recognition task common in NLP. It is also common to normalize temporal expressions after the extraction procedure.

A temporal expression (also *time expression*, or *timex*) is a sequence of tokens (words, numbers and characters) that can denote a point in time, duration or frequency. The concept of temporal expression is not strictly defined and indistinct. Nevertheless the ISO standard "*TimeML*" for labeling and normalization of temporal expressions was developed and adopted [10].

These are some examples of temporal expressions:

Что будут показывать <TIMEX>сегодня ночью</TIMEX> по пятому каналу? /What will be on TV tonight?

<TIMEX>8 сентября 2013 года</TIMEX> состоялись выборы на пост мэра Москвы. / Mayoral elections were held on the 8th of September 2013.

<TIMEX>Через 2 недели</TIMEX> состоится встреча с руководителем. / The meeting with the chief will be held in two weeks.

Какую телепередачу показывают <TIMEX>ежедневно в 7 часов вечера</TIMEX>? / What program is on TV every day at 7 p.m.

It should to be noticed that words like "*мгновенно*" ("*instantaneously*") or "*быстро*" ("*quickly*") are not temporal expressions.

There are two main approaches to solving the problem of temporal expressions extraction. The first one is a rule-based approach. The main idea of this approach is searching in sentence for predefined patterns of temporal expressions [9,12]. This approach requires developing a list of patterns of timexes composed by a linguist and the resulting list is usually dependent on the domain of texts [12]. The second approach is based on machine learning [9]. In the context of statistical approach linguistic knowledge is not necessary, but large amount of labeled data is needed for training a statistical model.

3.2. Labeling Scheme and Generation of Features

The "*TimeML*" specification suggests a XML-like markup for time expressions. But XML-like scheme is verbose and redundant for simple extraction of temporal expressions. For this reason we took simple and widely-used BIO (Begin Inside Out) and IO (Inside Out) labeling alphabets.

Now we describe how to reduce task of temporal expression extraction to task of sequence labeling.

1. If i^{th} token of input sequence is the first token of temporal expression, then the i^{th} output label is *B*.
2. If i^{th} token of input sequence is included in temporal expression and is not the first token of it, then then the i^{th} output label is *I*.
3. The output label is "O", otherwise.

The set of valid features of tokens *T* contains five types of features.

1. All possible analyses of the token retrieved from the dictionary of OpenCorpora (non disambiguated) [12]. Examples of such kind of features: "noun", "verb", "dative case", "perfective aspect", etc.

2. Features based on spelling of the token: “token starts with capital letter”, “token has a digits”, etc.
3. Features describing position of token in the sentence: “token is the first token in sentence”, “token is the last token in sentence”
4. Features indicated that token is a specific for temporal expression word, i. e. trigger-word: “token is a month name”, “token is a name of day of week”, etc.
5. Previous groups features of nearest tokens.

3.3. Dataset

Currently there is no dataset with labeled temporal expressions for Russian. For this reason a small set of Russian phrases was labeled by hands in BIO labeling scheme. This small dataset contained 2000 sentences with roughly 500 temporal expressions and was used like test dataset.

Nevertheless, training machine learning algorithms require much bigger dataset. So, we used semi automatic procedure based on regular expressions for obtaining training data. Moreover, we developed rule-based base-line algorithm with help of this regular expressions.

3.4. Feature Selection

Set of valid features T has a high dimension. So, it is reasonable to apply feature selection methods. We used algorithm “Random Forest” [4] as an algorithm for feature selection. This algorithm is related to stochastic logic methods of machine learning. But “Random Forest” is also applicable to classification problem. So we used it like alternative method of temporal expressions extraction. Advantages of “Random Forest” are high generalization ability, application to data with high dimension and ability to deal with binary features.

3.5. Experiments

Subset of dataset from project OpenCorpora was taken for training and testing. This subset was labeled automatically with base-line algorithm based on patterns. Then subset was split into train (380,000 phrases, 5,000 temporal expressions) and test parts (40,000 phrases, 9,000 temporal expressions). Moreover, the subset of sentences labeled manually was used for testing (2,000 phrases, 500 expressions).

We trained two alternative algorithms: “Random Forest” and CRF. Outputs of algorithms were converted from BIO scheme to IO. Then standard quality measures were calculated (Recall R, Precision P and F_1-measure):

- $P = \dfrac{tp}{tp + fn}$

- $R = \frac{tp}{tp + fp}$
- $F_1 = \frac{2PR}{P + R}$

The results of base-line algorithm, CRF and "Random forest" are listed in the table below. It should be noticed that these results were obtained on the best set of features, i. e. on features which were selected with feature selection procedure "Random Forest".

Table 1. Experimental results. Time expressions extraction

Algorithm	*P*	*R*	F_1
Base-line	95,7	85,7	90,4
RF	96,4	87,1	91,5
CRF	96,3	89,9	93,05

4. Shallow Parsing

4.1. Task Definition

The problem of shallow parsing was formultated in [1]. The shallow parser searches in text the so-called base NPs which are the fragments of noun phrases excluding recursive parts. The common example of base-NP in English is a noun with its left adjuncts:

green colorless thoughts,
USA President Barack Obama.

To adapt this approch to Russian we include in base NP agreed adjectives, numerals and nouns (*Президент Ельцин*), and dependent base NPs in genitive case. Thus, according to these criteria the following phrases may be considered as base NPs:

любимый руководитель Ким Чен Ир/beloved leader Kim Jong Il

*друг отца жены инженера Лаборатории эффективных алгоритмов/*a friend of father of the Algorithm Lab engineer's wife

and so on.

It should be noticed that base NPs (or *chunks*) do not have clear linguistic interpretation. Nevertheless, it was proposed in [1] that speakers tend to make pauses on the borders of base NPs. Thus base NP appears to be a psycholinguistic entity and correlate with the term *elementary discourse item* appearing in some linguistic theories.

4.2. Solution Outline

To reduce shallow parsing to the sequence labelling problem we took the following approach:

1. If the i^{th} input token is the beginning of a base NP then the i^{th} output label is B.
2. If the i^{th} input token is insid a base NP then the i^{th} output label is I.
3. Otherwise, the output label is O.

To provide the opportunity of detection of heads of the phrase we augmented the standard BIO-alphabet with two labels: BH (token is the beginning of the base NP and is the head) and IH (token is the phrase head). The input sequence of the shallow parser was the output of the morphological analyzer. The features were: part of speech, gender, number, case (if defined), upper/lowercase. We also used features of neighboring tokens and there combinations.

4.3. Dataset

The training set was generated from the syntactically annotated corpus SynTagRus IITP RAS [8]. Every syntactic tree corpus was traversed and subtrees with noun roots were detected. Then the following edges were excluded: 1) the edges coming into tokens different from nouns, adjectives, numerals or adverbs or to nouns in case different from genitive and non-agreed with the head; 2) the edges coming to non-neighboring tokens from the base NP. Based on these criteria *BIO-BH-IH* was generated for the training and test set. The training set comprised 40976 and 6310 were chosen for the test set.

4.4. Results

We trained two models: the first one was based only on the grammatical features and the second one also used input token (wordform) as a feature. The results are given in tables 2 and 3.

Table 2. Experimental results. Base NPs

Model	P	R	F_1
SynTagRus. Tokens+	93,39	93,07	93,23
SynTagRus. Tokens-	94,52	94,27	94,39
Pereira (2003)	n/a	n/a	94.38

Table 3. Experimental results. Heads detection

Model	P	R	F_1
SynTagRus. Tokens+	95,56	95,14	95,55
SynTagRus. Tokens-	96,48	95,45	95,96

The tables demonstrate the method can effectively detect both base NPs and the corresponding heads. For comparison we also give the results achieved by Pereira for English.

5. Noun Phrases Inflection

5.1. Task Definition

The problem of inflection of noun phrases i.e. changing its case from nominative to any oblique case generally can be solved by means of special phrase inflection rules. This approach however is time consuming and error prone, so it is reasonable to make use of machine learning approach. In fact, the task reduces to the search of the head and tokens agreed with it. Then, all found target tokens are set to the proper morphological form.

5.2. Solution Outline

Using the simplest possible label inventory consisting of “1” and “0” is sufficient in this case. We output label “1” of the token should be set in a target form and “0” otherwise. Composing the input sequence we considered only tokens with morphological features of nominals (e.g. noun, adjective, numeral, participle). Neighboring non-nominal tokens were used as features. Thus, each input token had the following features: part of speech, gender, number, case and features of neighboring tokens both included and non-included in the input sequence.

For example, for the NP *медведь из леса* (a bear from forest), token *леса* would have following features: *noun, masculine, genitive, preposition_in_posistion-1, noun_in_position_-2*. We also used combination of the features.

5.3. Dataset

Corpus of noun phrases was generated from syntactic trees of SynTagRus corpus. We had to use information about edge types in the tree to generate cleaner training set. Each generated phrase was set to nominative case. We used noun phrases of the length no more than 10 tokens. Unfortunately, although we used edge labels the corpus still contained many erroneous entries. We manually cleaned a corpus of 100,000 tokens. 10,000 form them were selected for the training set.

5.4. Results

We present the results of experiments on the search of targets of inflection:

Table 4. Experimental results. Inflection targets

P	R	F_1
99,44	99,74	99,59

Although we anticipated good results the algorithm happened to perform better than expected and the total time spent on the development less than the estimated time on the grammar preparation.

6. Conclusions

Solution of three natural language processing tasks including shallow parsing, temporal expressions extraction and noun phrase inflection have been proposed. We have shown that all these tasks can be reduced to sequence labeling problems and solved by means of linear-chain CRF statistical model. It allows replacing complex linguistic rules with a set of relevant features and data preparation. This work is often less time consuming and error prone.

References

1. *Abney S.* (1991), Parsing by chunks, Principle-based Parsing, Kluwer Academic Publishers, pp. 257–279.
2. *Antonova A. Ju., Solovyev A. N.* (2013), Conditional random field models for the processing of Russian [Ispol'zovanie metoda uslovnyh sluchajnyh polej dlja obrabotki tekstov na russkom jazyke], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2013" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2013"], Bekasovo, pp. 39–52.
3. *Bottou L.* (1991), A theoretical approach of connectionist learning: Applications to the recognition of speech [Une approche theorique de l'apprentissage connexionniste: Applications a la reconnaissance de la parole], Doctoral dissertation [Doctoral dissertation], University of Paris XI.
4. *Breiman L.* (2001), Random forests, Machine Learning, Vol. 45, no. 1, pp. 5–32.
5. *Culotta A., Wick M., Hall R.* (2007), First-Order Probabilistic Models for Coreference Resolution, In Proceedings of HLT/NAACL, pp. 81–88.
6. *Granovskij D. V., Bocharov V. V., Bichineva S. V.* (2010), Open corpora: principles of work and prospects [Otkrytyj korpus: printsypy raboty i perspektivy], In proceddings of XIII Russian Conference "Internet and society today" [Internet i sovremennoe obshchestvo: Trudy 18 Vserossijskoj ob"edinennoj konferentsii], St. Petersburg, pp. 94–99.
7. *Lafferty J., McCallum A., Pereira F.* (2001), Conditional random fields: Probabilistic models for segmenting and labeling sequence data, Proceedings of the 18th International Conference on Machine Learning, Williamstown, Massachusetts, pp. 282–289.

8. *Nivre J., Boguslavsky M., Iomdin L.* (2008), Parsing the SynTagRus treebank of Russian, Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pp. 641–648.
9. *Poveda J., Surdeanu M., Turmo J.* (2007), A comparison of statistical and rule-induction learners for automatic tagging of time expressions in english, In Proceedings of the 14th International Symposium on Temporal Representation and Reasoning (TIME 2007), IEEE, pp. 141–149.
10. *Pustejovsky J., Ingria R., Sauri R.* et al. (2003), Timeml: Robust specification of event and temporal expressions in text, In Fifth International Workshop on Computational Semantics (IWCS-5).
11. *Ramshaw L. A., Marcus M. P.* (1995), Text chunking using transformation-based learning, The Third Workshop on Very Large Corpora, pp. 82–94.
12. *Reeves R. M., Ong F. R., Matheny M. E.* (2013), Detecting temporal expressions in medical narratives, I. J. Medical Informatics, Vol. 82, no. 2, pp. 118–127.
13. *Sha F., Pereira F.* (2003), Shallow parsing with conditional random fields, In Proceedings of HLT/NAACL, pp. 213–220.
14. *Sutton C., McCallum A.* (2006), An Introduction to Conditional Random Fields for Relational Learning, MIT Press.

КОНСТРУКЦИИ С СОЮЗОМ *ЧТОБЫ*: РЕСУРСЫ И СООТВЕТСТВИЯ[1]

Кустова Г. И. (galinak03@gmail.com)

Институт русского языка
им. В. В. Виноградова РАН, Москва, Россия

В статье рассматриваются сложные предложения с существительным в главной части и союзом *чтобы* в придаточной. Конструкция желательного признака (1а: *Где еще найдешь сиделку, чтобы хорошо ладила с Васей?*) и конструкция функционального несоответствия (1б: *Он не дама, чтобы ему цветы дарить*) сравниваются с ресурсной конструкцией (2: *У нас есть время, чтобы сходить в кино*).

Ключевые слова: конструкция, союз, ирреальное следствие

CONSTRUCTIONS WITH THE CONJUNCTION *CHTOBY*: RESOURCES AND CORRELATIONS

Kustova G. I. (galinak03@gmail.com)

Vinogradov Russian Language Institute
of the Russian Academy of Sciences, Moscow, Russia

The article deals with complex sentences with a noun in the main clause and the conjunction *chtoby* in the subordinate clause. Construction «desirable feature» (*Gde najdesh sidelku, chtoby xorosho ladila s Vasej?*) and construction «functional inconsistency» (*On ne dama, chtoby emu cvety darit'*) are compared with the resource construction (*U nas est' vremy'a, chtoby sxodit' v kino*).

Key words: Construction, conjunction, unreal consequence

[1] Работа выполняется при поддержке РГНФ, проект № 14-04-00507а.

Есть целое семейство сложных предложений, у которых в главной части — существительное, а в придаточной — союз *чтобы*[2]:

(1) а. *Где сейчас найдешь няньку, чтобы ладила и с ребенком, и с родителями?*
Есть поблизости ресторан, чтобы недорого и тихо?
б. *Я не попугай, чтобы повторять чужие слова*
Я не банк, чтобы всем деньги раздавать
Здесь не юг, чтобы абрикосы
Сейчас не лето, чтобы без пальто ходить
в. *Не те времена, чтобы арестов бояться / Не тот человек, чтобы перед трудностями отступать / Не та погода, чтобы без пальто ходить / Не тот возраст, чтобы балетом заниматься*
г. *Не надо быть детективом, чтобы догадаться*

Эти предложения различаются по структуре и свойствам, и в них можно выделить несколько типов, но для каждого из этих типов можно найти довольно близкие аналогии с хорошо известными и описанными в грамматиках моделями, где встречается ирреальное следствие:

- во-первых, это ресурсные конструкции со значением достаточности / недостаточности или наличия / отсутствия ресурса: *У нас достаточно времени, чтобы собраться*; *У партии есть ресурсы, чтобы идти на выборы*; *У нас нет комнаты, чтобы поселить гостей*;
- во-вторых, это так называемые местоименно-союзные предложения (в главной части — местоимение (с отрицанием), в придаточной — ирреальное следствие): *Здесь не так тепло, чтобы абрикосы сажать*; *Не такой он человек, чтобы отступать перед трудностями*;
- в-третьих, фразеологизованные конструкции со значением следствия ([АГ-80, т. 2, § 3077]): *Откос был слишком крутой, чтоб удержаться на нем* (пример из [АГ-80]).

Таким образом, предложения типа (1) с существительным в главной части и союзом *чтобы* легко соотносятся с существующими рубриками синтаксической классификации. Тем не менее сами эти предложения в грамматиках не упоминаются: отсутствует не только их описание, но даже перечень таких конструкций.

Есть отдельные упоминания таких или похожих предложений в разных источниках. Например, в учебнике по синтаксису [Крючков, Максимов 1977] среди фразеологизованных моделей придаточных цели есть модель: *Я не больной, чтобы сидеть дома / Разве я больной, чтобы сидеть дома* (экспрессивное отрицание + ирреальное следствие). В работе [Крейдлин 1992] рассмотрена конструкция: *На то и учитель, чтобы учить* — тоже включающая существительное, однако принципиальным (и непосредственно коррелирующим со *чтобы*-придаточным) в этой конструкции является элемент *на то*, который генетически связан с предназначением (статья Г. Е. Крейдлина посвящена

[2] Литературные примеры извлечены из Национального корпуса русского языка; некоторые из них приводятся в сокращении.

семантике предназначения, близко связанной с семантикой цели); конструкция *на то и… чтобы* упоминается также в [АГ-80] среди фразеологизованных целевых конструкций с несобственно целевым значением (т. 2, § 3060). Наконец, в [АГ-70] приводятся предложения вида *На всяком деле должен быть крепкий человек, чтобы его слушались без крику*; *Нет человека, чтобы ему нечего было о себе рассказать* — но они считаются все-таки местоименно-союзными с незамещенной позицией местоимения (ср.: *должен быть такой человек, чтобы…*; *нет такого человека, чтобы…*).

Для предложений типа (1) важен целый ряд параметров — семантика глагола в главной части (например, бытийная, но не только), модальность, отрицание и т. д. Но мы рассмотрим лишь одну их особенность — контаминированный характер, делающий их переходными структурами, не вписывающимися в стандартные классификации.

Аппарат традиционной грамматики не располагает такими средствами, чтобы включить предложения (1), построенные на базе существительного и союза *чтобы*, в какой-либо известный тип. Собственно, возможных типов, связанных с существительным, два: атрибутивные предложения, где придаточные прикрепляются к существительному, и местоименно-союзные, где придаточные прикрепляются к сочетанию местоимения *такой* с существительным.

К атрибутивным предложения (1) отнести нельзя, т. к. в атрибутивных предложениях реализуется анафорическая связь относительного местоимения и существительного, поэтому в них невозможен союз: *лектор, которого* (= *лектора*) *мы пригласили*. При этом очевидно, что, например, предложения *Где взять помощника, который бы знал два языка* и *Где взять помощника, чтобы знал два языка* по смыслу практически не отличаются.

Второй тип — местоименно-союзные предложения — требует наличия обязательного местоимения, с которым и соотносится придаточное: *Он не такой человек, чтобы отказать другу в помощи*.

В примерах (1) есть такие, куда местоимение легко вставляется, и даже такие, где местоимение присутствует, но другое — не *такой*, а *тот*, ср. (1в). Однако, во-первых, это не значит, что такие предложения — редуцированная местоименная модель. Вообще, вопрос об обязательности / необязательности местоимений в структуре сложного предложения очень непростой, и мы не можем его подробно рассматривать. Заметим, однако, что, например, предложения вида *Он побледнел, как будто увидел привидение* не считаются в грамматиках редуцированным вариантом предложений вида *Он так побледнел, как будто увидел привидение* — такие предложения грамматики относят к разным типам: местоименно-союзному и сравнительному соответственно. А во-вторых, в группе (1) есть такие предложения, куда при всем желании местоимение вставить нельзя: *Я не прокурор, чтобы кого-то обвинять*.

В данной работе мы рассмотрим только два типа из группы предложений (1)–(1а) (*Есть поблизости ресторан, чтобы недорого и тихо?*) и (1б) (*Я не попугай, чтобы повторять чужие слова*); соответствующие им конструкции будем обозначать квадратными скобками — [1а] и [1б]. Чтобы показать специфику этих моделей, мы будем рассматривать их на фоне ресурсной конструкции [2]

(*У нас есть деньги, чтобы поехать в отпуск*). Все три конструкции имеют нечто общее, а именно — они выражают соответствие (или несоответствие): [2] — соответствие возможностей (ресурсов) субъекта его целям, [1а] — соответствие предмета требованиям (запросам, желаниям) субъекта, [1б] — несоответствие реальных свойств предмета чьим-то ожиданиям. При этом конструкции [1а] и [1б], несмотря на некоторое сходство, довольно существенно отличаются по своим свойствам как друг от друга, так и от ресурсной конструкции [2].

Ресурсная конструкция

Под ресурсной конструкцией мы будем понимать сложное предложение, в главной части которого упоминается ресурс, нужный для достижения некоторой цели.

Общая схема ресурсной конструкции такова:

«Ресурс R, имеющийся у субъекта S, позволяет субъекту S реализовать цель P»

Конструкция имеет множество вариантов (моделей):

(2) *Есть время, чтобы собраться*
Нет времени, чтобы собраться
Нужно время, чтобы собраться
Достаточно времени, чтобы собраться
Хватает времени, чтобы собраться
Требуется время, чтобы собраться

ПРИМЕЧАНИЕ. Вторая часть конструкции может выступать в форме инфинитива: *Нет времени зайти в магазин*, но мы эту возможность не рассматриваем.

Многие из этих моделей описаны в лингвистических работах, см. особенно [Левонтина 2006: 219–220], где конструкции типа *Мне нужно полчаса, чтобы доехать до работы* рассматриваются как потенциальная цель, и [Богуславский 1996: 97–100 и др.], где описаны валентности «несущего» элемента ресурсных конструкций — слов типа *достаточно*, *хватает*, *нужно*. Ресурсная конструкция упоминается (конечно, не под таким названием) также в [АГ-80, т. 2, § 3060], где она включена в раздел целевых предложений с несобственно целевым значением.

Ресурсная конструкция обычно считается модификацией целевых предложений. И она действительно тесно связана с целевой:

- в целевой главная часть — действие субъекта, придаточная часть — его цель;
- в ресурсной в главной части — наличие, необходимость, достаточность, недостаток средства (ресурса) у субъекта-обладателя, в придаточной — цель, т. е. ситуация, желательная / запланированная для обладателя средства (в работе [Левонтина 2006] справедливо отмечается, что такая цель

не обязательно есть у субъекта, но если бы он ее поставил, то для ее реализации потребовался бы указанный ресурс).

В плане дальнейшего изложения для нас существенно, что (а) в ресурсной конструкции есть субъект-обладатель ресурса; он может быть выражен косвенным падежом (*у S-а есть … / у S-а достаточно …*) или вообще не выражен синтаксически; но мыслится использование ресурса субъектом-обладателем при реализации Р; (б) ситуация Р будет реализовываться именно субъектом S (не случайно предикат придаточного выражен инфинитивом, как бывает в целевых придаточных при кореферентности субъекта).

Т. е. в обеих частях ресурсной конструкции один и тот же субъект S — это его ресурс и его будущее — возможное / желательное / запланированное — действие Р. Это важно для сравнения с конструкцией [1а].

Для ресурсной конструкции отрицание не существенно, т. е. не является конструктивным элементом; ресурсная конструкция может иметь и положительную, и отрицательную форму:

У нас есть время, чтобы зайти в магазин vs. *У нас нет времени, чтобы зайти в магазин*. Это важно для сравнения с конструкцией [1б].

Конструкция [1а]: ПОТРЕБНОСТЬ в ОБЪЕКТЕ с ЗАДАННЫМ / ЖЕЛАТЕЛЬНЫМ ПРИЗНАКОМ

Для краткости будем называть [1а] конструкцией желательного признака. Она похожа на ресурсную, но устроена иначе:

А Анастасия Владимировна терпела: где еще найдешь сиделку, чтобы хорошо ладила с Васей? [Александр Терехов. Каменный мост (1997–2008)]

В главной части речь идет о потребности субъекта S в объекте X (о поиске объекта X) с заданным признаком, а во второй — описывается сам этот признак Р.

Существование объекта X — в сфере действия модального оператора, а сам объект нереферентный, ср.: **Наконец нашли сиделку, чтобы ладила с Васей*.

Ситуация Р — это одновременно признак объекта и потребность, желание субъекта, т. е. этот признак (объект с таким признаком) соответствует потребностям субъекта. От ресурсной конструкции такие предложения отличаются тем, что ситуацию Р может реализовать (или будет реализовывать) не субъект потребности S, а объект X (упоминаемый в главном предложении в контексте поиска, потребности, желательности):

Попробуйте найти умельца [X], *чтобы он* [X] *собрал все листы в стопку, по корешку сделал штук 15 пропилов ножовкой (неглубоко, 3–4 мм) и в пропилы вклеил шнуры, лучше льняные*. [коллективный. Реставрация книг. Переплетное дело (2009)]

Лишь бы что-нибудь купить, найти мужика, чтобы за нее заплатил, свозил бы ее на курорт… [«Хулиган», 2004.08.15]

— Неужели командование не нашло офицера, чтобы составить компанию дочери начальника штаба армии? [Вадим Кожевников. Щит и меч. (1968)]

Когда он репетировал, к нему подошли и спросили: — Вам нужен помощник, чтобы страницы переворачивать? [Юрий Башмет. Вокзал мечты (2003)]

Т. е. важным отличием конструкции [1а] от ресурсной является то, что в ситуации P заинтересован один субъект (S), а реализовывать ее будет другой субъект — X (в ресурсной же конструкции ресурс нужен субъекту S для того, чтобы самому реализовать P).

Ситуация P не обязательно является действием X-а, а может быть просто его признаком:

Он идет по зеленому лугу, Смачно хрюкает и ревет — Он желает найти подругу, Чтобы тоже была бегемот. [Никита Богословский. Заметки на полях шляпы (1997)]

Наиболее распространенный случай (из примеров, найденных в НКРЯ) — предложения, где X — лицо. Но X может быть и местом — функциональным (учреждением), природным объектом или условным пространством:

Есть ресторан, чтобы недорого и тихо?

— Где бы найти место, чтобы уважали умного, — вздохнул чех.
[Андрей Лазарчук, Михаил Успенский. Посмотри в глаза чудовищ (1996)];

Но было очень трудно найти среду, чтобы она подчинялась законам механики, а в ней распространялись упругие волны, которые можно отождествить с волнами электромагнитными. [«Наука и жизнь», 2006]

Важно, чтобы в придаточном был именно признак X-а, а не действие субъекта S, иначе получится ресурсная конструкция или просто придаточное цели, ср.:

Я всегда найду льдину, чтобы взлететь на «У-2».
[Л. К. Бронтман. Дневники и письма (1943–1946)]

Однако такую целевую конструкцию легко преобразовать в конструкцию желательного признака, ср.:

Я всегда найду льдину, чтобы не раскололась при взлете.

Такие конструкции близки к атрибутивным придаточным с сослагательным наклонением или будущим временем (которое ведет себя как модальность):

Я всегда найду льдину, которая бы не раскололась при взлете / которая не расколется при взлете.

Существительное с предметным значением в главной части тоже возможно — с тем же условием — в придаточном должно быть желательное свойство предмета, а не запланированное действие субъекта:

Мне нужна таблетка, чтобы растворялась в воде;
Где же взять устройство, чтобы помещалось в кармане.

На периферии конструкций [1а] — предложения, в которых признак P, выраженный в придаточном, не является желательным для конкретного субъекта, а просто обсуждается существование (или несуществование) объекта с таким признаком:

Иначе род людской должен был бы непременно прекратиться: где найти идиотку, чтобы стала рожать, помня о том, что было с ней в прошлый раз… [Андрей Волос. Недвижимость (2000) // «Новый Мир», 2001]

Итак, конструкция желательного признака — своего рода контаминация атрибутивной конструкции, описывающей признак предмета, и целевой конструкции, описывающей цель (желание) человека:

Есть ресторан, чтобы тихо и недорого? — 'S хочет, чтобы ресторан был тихий и недорогой, и интересуется, есть ли такой ресторан / ищет такой ресторан'.

У *чтобы*-придаточного как бы двойная семантическая связь: с существительным в главной части, признак которого обозначен в придаточном, и с — часто невыраженным, но присутствующим в семантической структуре — субъектом S, для которого этот признак P является желательным.

Конструкция [1б]: ФУНКЦИОНАЛЬНОЕ НЕСООТВЕТСТВИЕ

В конструкции [1б] существительные в главной части могут иметь разную семантику:

- лицо:

Я не прокурор, чтобы кого-то обвинять
Я не повар, чтобы на всех готовить

Я не сторож, чтобы ваше добро охранять
Я не прислуга, чтобы за вами убирать

- место:

Здесь не пляж, чтобы в купальнике ходить
У нас не юг, чтобы абрикосы сажать
Тут не Сочи, чтобы пальмы росли
Тут не ресторан, чтобы блюда выбирать

- время:

Сейчас не зима, чтобы на лыжах кататься
Сейчас не лето, чтобы без пальто ходить

ПРИМЕЧАНИЕ. Особую группу в данной модели образуют предложения с параметрическими существительными с причинным значением (логического обоснования):

— Подумаешь! Не причина, чтобы травиться. Мог бы объявить себя банкротом. [Леонид Юзефович. Дом свиданий (2001)]

— Это не повод, чтобы бросать работу, за которую платят приличные деньги! [Максим Милованов. Кафе «Зоопарк» (2000)]

Эмоции — это не резон, чтобы мы теряли бешеные деньги. [Р. Б. Гуль. Азеф (1958)]

Но такие предложения мы не рассматриваем.

Конструкции [1б] имеют разные варианты (или модификации, или аналоги), в которых существительное стоит не в именительном падеже или даже вообще отсутствует:

Вы не на пляже, чтобы в купальниках ходить
Это же было не вчера, чтобы я все помнил
Я же не каждый день кодекс читаю, чтобы помнить все статьи

Но человечество ещё не дозрело, чтобы эту шифровку прочитать. [Людмила Улицкая. Казус Кукоцкого
[Путешествие в седьмую сторону света] // «Новый Мир», 2000]

Психика у него еще не сформирована, чтобы выдержать то, что выдерживает сегодня Кафельников.
[Шамиль Тарпищев. Самый долгий матч (1999)]

Он [человек] *еще не дорос, чтобы пробиться к другим планетам, через хаос неорганизованной материи космоса*. [И. А. Ефремов. Лезвие бритвы (1959–1963)]

Часто такие предложения легко преобразовать в конструкции с существительным или прилагательным:

Где у меня деньги, чтобы всем раздавать!

Разве я деньги лопатой гребу, чтобы всем раздавать! — ср.: *Я не банк / не банкир, чтобы всем деньги раздавать.*

Коммунизм еще не наступил, чтобы я совершал такие поступки. [Е. Попов. Мыслящий тростник (1970–2000)] — ср.: *Еще не коммунизм, чтобы я совершал такие поступки*.

Еще не выросла, чтобы со мной на таком тоне разговаривать. [В. Распутин. Деньги для Марии (1967)] — ср.: *Мала еще, чтобы со мной так разговаривать*.

Но мы ограничимся только «субстантивными» предложениями.

Очевидно главное различие [1б] с предложениями заданного признака [1а]: в [1б] существительное находится в позиции сказуемого — т. е. это не «предмет», а предикат, признак. Существительное в главной части актуализируется признак F (часто это род занятий, но не обязательно), причем для лица с таким признаком F действие P, обозначенное в придаточном, является типичным, ожидаемым или даже обязательным (если предусмотрено профессиональными обязанностями): прокурор обвиняет, повар готовит, сторож охраняет и т. д.; вместо функционального признака может быть представлено имя собственное, субъект которого ассоциируется с каким-либо характерным проявлением, ср.: *Я не Пушкин, чтобы стихи писать*. Действие P может быть типично для некоторого лица в месте F (на пляже ходят в купальниках), во время F (летом ходят без пальто), т. е. P является импликацией F. При этом конструкция [1б] сообщает, что лицо X не обладает признаком F, и от него не следует ждать действия P.

Эта конструкция, в отличие от предыдущей, очевидным образом экспрессивная и полемическая. Для импликативной конструкции отрицание принципиально, причем это экспрессивное отрицание — т. е. либо собственно отрицательная конструкция, либо риторический вопрос:

Я не прокурор, чтобы кого-то обвинять!
Разве я прокурор, чтобы кого-то обвинять!

Предложения [1б] употребляются в ситуации, когда некоторый наблюдатель (собеседник) S ожидает, или предполагает, или допускает со стороны

X-а действия P. Конструкция сообщает, что ожидания наблюдателя являются необоснованными:

Я не пожарный, чтобы лезть на крышу по первому требованию.
[И. А. Бунин. Шаляпин (1938)] — ‘не стоит ожидать P’.

Возможны и другие модальности:

Здесь не пляж, чтобы в купальниках ходить — ‘не следует делать P’.

Интересно, однако, что у [1б] есть другой вариант, где то же существительное является не сказуемым (ср. *Я не прислуга, чтобы за вами убирать*), а подлежащим — обычно при бытийном глаголе:

(1б’) *Есть прислуга, чтобы за вами убирать, а я секретарь*.

Эта конструкция тоже полемическая, произносится с особой интонацией и обычно подразумевает (или имеет) продолжение: *Есть прислуга, чтобы P, — а я не прислуга*.

Формально [1б’] похожа на ресурсную конструкцию, но от ресурсной конструкции она отличается тем, что в ресурсной конструкции *есть* употребляется в значении ‘иметь’, а не ‘существовать’, соответственно, в семантической структуре присутствует субъект обладания, даже если синтаксически он не выражен:

Есть повар, чтобы готовить, есть сторож, чтобы территорию охранять, есть прислуга, чтобы убирать, — что еще надо?

Таким образом, в [1б], как и в [1а], тоже обнаруживается двойная связь придаточного: с одной стороны, P актуализируется наблюдателем, который ожидает, предполагает или желает действия P со стороны X-а [иначе бы нечего было опровергать], с другой, P — это импликация признака F, типичная функция, типичное проявление для лица, места, времени с признаком F.

Понятно, почему конструкции типа (1) не охватываются и не учитываются традиционными грамматическими описаниями: они являются своего рода контаминациями, переходными, промежуточными образованиями, соединяющими признаки разных структур, а такие образования всегда представляют трудности для классификации, сопротивляются попыткам поместить их в ту или иную «клетку».

Однако, с другой стороны, они представляют и очевидный интерес, т. к. позволяют проследить живые связи между разными участками языковой системы, процессы переходов, синонимические механизмы выражения смысла, действующие в языке.

Литература

1. *АГ-70* — Грамматика современного русского языка. М., 1970.
2. *АГ-80* — Русская грамматика, т. 2, М., 1980.
3. *Богуславский И. М.* Сфера действия лексических единиц. М., 1996.
4. *Крейдлин Г. Е.* К проблеме языкового анализа концептов «цель» vs. «предназначение» // Логический анализ языка. Модели действия. М., 1992. С. 23–30.
5. *Крючков С. Е., Максимов Л. Ю.* Современный русский язык. Синтаксис сложного предложения. М., 1977.
6. *Левонтина И. Б.* Понятие цели и семантика целевых слов русского языка // в кн.: В. Ю. Апресян, Ю. Д. Апресян, Е. Э. Бабаева, О. Ю. Богуславская, Б. Л. Иомдин, Т. В. Крылова, И. Б. Левонтина, А. В. Санников, Е. В. Урысон. Языковая картина мира и системная лексикография. Отв. ред. Ю. Д. Апресян. М., 2006. С. 163–238.

References

1. *AG-70* — Grammatika sovremennogo russkogo jazyka. M., 1970.
2. *AG-80* — Russkaja grammatika. M., 1980.
3. *Boguslavsky I. M.* Sfera dejstvija lexicheskih edinic. M., 1996.
4. *Krejdlin G. E.* K problem jazykovogo analiza konceptov «cel'» vs. «prednaznachenije» // Logicheskij analiz jazyka. Modeli dejstvija. M., 1992. P. 23–30.
5. *Kr'uchkov S. E., Maximov L. Ju.* Sovremennyj russkij jazyk. Sintaksis slozhnogo predlozhenija. M., 1977.
6. *Levontina I. B.* Pon'atie celi i semantika celevyh slov // Apresjan Ju. D. i dr. Jazykovaja kartina mira i sistemnaja lexikografija. M., 2006, pp. 163–238.

ОПИСАНИЕ ЛОКАТИВНЫХ ЗАВИСИМЫХ В СИСТЕМЕ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Леонтьев А. П. (Aleksey_L@abbyy.com),
Петрова М. А. (Maria_Pet@abbyy.com)

ABBYY, Москва, Россия

THE DESCRIPTION OF LOCATIVE DEPENDENCIES IN A NATURAL LANGUAGE PROCESSING MODEL

Leontiev A. P. (Aleksey_L@abbyy.com),
Petrova M. A. (Maria_Pet@abbyy.com)

ABBYY, Moscow, Russia

The paper suggests semantic and syntactic descriptions of locative dependencies in an NLP model and focuses on the problems which locative adjuncts evoke for a system aimed at different tasks based on semantic analysis, especially at machine translation. A formal description of locative groups faces several problems. The first is the definition of locative semantic relations between words, as locative dependencies can have different meanings, such as the meanings of initial and final points (*walk [from/to the door]*), route (*walk [across the room]*), and others. Second, one has to define the set of words that can fill locative adjuncts, and the border between the locative and non-locative groups is not always distinct: *in the street* is definitely a locative, but what about *on the Internet* or *in a meeting*?
Third, the syntactic realizations of locative senses are rather numerous. On the one hand, locative adjuncts include many prepositions with different semantics—like *on, in, under, above*, etc. On the other hand, different nouns combine with different prepositions to denote the same meaning, like *in the country*, but *on the island*.
The current paper suggests a formal approach appropriate for dealing with all these difficulties.

Keywords: semantics, syntax, NLP, machine translation, locative dependencies

In memory of Taisia Shtaryova

0. Introduction

The current paper is devoted to the formal description of locative dependencies in the ABBYY Compreno model—an integral NLP model, aimed at the semantic analysis of texts in natural languages.

Presentation of space meanings in natural languages has already been widely discussed in linguistics: there are works that focus on the conceptualization and categorization of space relations in a language and works devoted to the description of lexical and grammatical units that express locative meanings, especially prepositional constructions with space semantics and predicates with the meanings of position and motion.

Locative prepositions as well as different aspects of the space category have been analyzed in the works of A. V. Bondarko, V. G. Gak ([Gak 1996, 2000]), S. Feigenbaum, D. Kurzon ([Feigenbaum, Kurzon 2002]), M. V. Vsevolodova, E. Ju. Vladimirskij ([Vsevolodova 2010; Vsevolodova, Vladimirskij 1982]), V. A. Plungian, M. V. Filipenko, D. Paillard, O. N. Seliverstova (for instance, see [Paillard, Seliverstova (eds.) 2000]), L. Talmy ([Talmy 1983]), N. Ju. Shvedova ([Shvedova 1980]), and many others.

In addition, locative groups have been much discussed in modern works on cognitive linguistics, for example in works by [Aurnague et al. 2007, Bloom et al.1996, Hickmann, Robert 2006, Levinson, Wilkins 2006, Levinson 2003, Shay, Seibert 2003, Svorou 1994, van der Zee, Slack 2003].

In this article we want to focus on the problems of dealing with locative dependencies that arise when creating an NLP model aimed at machine translation, semantic search and other tasks based on the semantic analysis of texts, as formal description of locative groups faces both semantic and syntactic difficulties.

First, one should define the set of locative semantic relations between words, as the domain of groups with space semantics includes not only groups like *on the shelf/in the street*, but also groups with the meanings of initial and final point like *walk [from/to the door]*, route (as in *walk [across the room]*), and others. Second, it is necessary to define the set of words that can be used as locatives, as it is not always clear where to make the border between locative and non-locative adjuncts, namely, whether to regard as locatives only groups like *on the table*, or groups like *on/in the Internet* and *in a meeting* as well.

Third, even locative groups like *on the shelf* may cause difficulties, as these adjuncts include quite a large number of prepositions with different semantics: *on, in, under, above*, etc., and, in addition, different nouns are combined with different prepositions to denote the same meaning, like *in the country*, but *on the island*. All these things can be problematic for a formal model, especially for the task of machine translation.

The structure of the paper is as follows: the first part gives a short description of the general principles of the ABBYY Compreno formalism that are necessary for

further discussion. The second part is devoted to the semantic part of the locative description in the Compreno model and to the problems that arise within the semantic pattern of the formalism. The third section presents the syntactic part of the description, and the conclusion summarizes the results.

1. The ABBYY Compreno Formalism

The ABBYY Compreno formalism is an NLP model, which consists of several patterns: morphological, syntactic, semantic and statistical. All of them were presented in general terms in [Anisimovich et al. 2012] and, in addition, detailed description of the semantic pattern is provided in [Manicheva et al. 2012, Petrova 2014], and the syntactic pattern is presented in [Bogdanov, Leontyev 2013, Zuyev et al. 2013]. Now the descriptions of English and Russian are available, so here we will restrict ourselves to the English and Russian material.

Lexical content is organized in the form of a thesaurus-like semantic hierarchy (SH) with a tree structure, which consists of semantic classes (SC)—language-independent nodes that are filled with lexical units in natural languages. For instance, a SC BOY is filled with *boy* in the English part of the SH and, correspondingly, *mal'chik* in the Russian part. Currently the number of universal SCs is more than 110,000. The English part of the SH includes about 130,000 English notions, and the Russian part—about 120,000. The lower the SC, the less general notion it expresses, thus the ancestors of BOY are CHILD > PERSON_BY_AGE > … > HUMAN > BEING > PHYSICAL_OBJECT.

All classes are provided with additional semantic information through semantemes (marked with a <<>> sign): for instance, people and animals have a semanteme <<Animate>>, while food and plants have a semanteme <<Eatable>> (for more information on semantemes see [Anisimovich et al. 2012]).

Semantic links between words are presented through semantic slots, or deep slots (DS): the notion of DS is close to the notion of semantic valency, but the difference is that valencies are usually associated with actant slots only, like Agent or Object, whereas under DS we understand any dependency a word can attach. For example, the verb *to walk* has an [Agent] DS for *boy* in *[the boy] walks*, [Locative_FinalPoint] for *home* in *the boy walks [home]*, and [Ch_Parameter_Speed] for *fast* in *the boy walks [fast]*. Each slot can be filled with a strict set of SCs, e.g., [Agent] is filled with classes denoting beings, organizations and countries. The model includes more than 300 DSs in total.

Syntactically, DSs are expressed through language-specific surface slots (SurfS), thus [Agent] in *[the boy] walks* is expressed through the $Subject slot, or [Agent] in *the work is done [by the boy]* is expressed through $Object_Indirect_By (SurfSs are marked with a $ sign). A more detailed description of SurfSs is provided in section 3.

There are valencies for which the description of both DSs and SurfSs causes significant problems, and the locative groups are a good example here.

Let us now consider the description of the locative dependencies in the present model and discuss the difficulties that occurred while describing them.

2. The Semantic Description of Locative Dependencies in the ABBYY Compreno Model

There are two main problems for semantic description of locatives in the formalism described above.

First, one has to define the set of locative semantic relations, as there are locative valencies with different semantics, so we introduce not only the [Locative] slot, but slots like [Locative_InitialPoint] (for the locative of initial point), [Locative_FinalPoint] (for the locative of final point), and others as well.

Second, we have to restrict the filling of the slots with the classes of locative semantics only, and this introduces the question of where one should define the border between the locatives and non-locatives: *in the street* is definitely a locative, but how should one treat *on the Internet* or *in a meeting*? On the one hand, such groups are close to the locative groups both in their semantics and in their syntactic realization, yet on the other hand, they do not denote prototypical places, so we introduce additional locative slots with similar meaning but with different filling, for example, [Metaphoric_Locative] and [Locative_Event].

According to these parameters—the semantics of the locative valency and the slot's filling—we introduce several classes of locative DSs. In the current paper we suggest a description of three classes, namely, the [Locative_Class], the [Locative_InitialPoint_Class] and the [Locative_FinalPoint_Class], which are presented in Figure 1.

[Locative_Class]	**[Locative_InitialPoint_Class]**	**[Locative_FinalPoint_Class]**
[Locative]	[Locative_InitialPoint]	[Locative_FinalPoint]
[Metaphoric_Locative]	[Metaphoric_InitialPoint]	[Metaphoric_FinalPoint]
[Locative_Event]	[LocativeEvent_InitialPoint]	[LocativeEvent_FinalPoint]

Fig. 1. Classes of locative semantic slots

[Locative] is filled with entities that can denote places in a wide sense, namely, with countries, regions, spaces, physical objects, or beings. Thus, in example (1), all constituents in square brackets can be analyzed as [Locative]:

(1) *Its priority tasks today include building relations with the compatriots* [abroad].
Your assistant is still holding 10 cards [under the table].
Only 13 rodent species are found [on the island].

As the necessary fillers are positioned in different places of the SH, it is convenient to use a distributional semanteme here to mark them—the semanteme <<Place>> has been introduced for this purpose. So the filling of the [Locative] slot does not have to include a large number of small SCs, instead, we can indicate several branches of higher levels of the SH and restrict them with the <<Place>> semanteme. For example, the ENTITY class consists of different descendants: physical objects (like *table* or *bag*), mental objects (like *idea*, *thought*, or *opinion*), abstract and scientific objects (like *formula*

or *logarithm*), countries, organizations, and so on. It is clear that not all of them can be used as locatives, so we mark with the <<Place>> semanteme only the 'locative' classes.

[Locative_InitialPoint] and [Locative_FinalPoint] are filled with the same set of SCs and denote locatives of initial and final point, respectively, as in examples (2) for the former and (3) for the latter:

(2) *Don't forget that you can shoot* [from the side].
The witness saw this [from the window of her home].

(3) *You have come* [home].
Alice is coming [to the palace of the queen].

Syntactically, all these slots usually correspond to prepositional nominal groups, which include all possible locative prepositions, and adverbs, so the syntactic realizations of the slots can be rather numerous, as one can see from the examples in (1–3).

[Metaphoric_Locative] is filled with words like *imagination*, *book*, *Internet*, *head*, *TV*, or *document* that do not denote physical space in a literal sense, but are close to the locative groups both in their semantics and in syntactic realizations, see the examples in (4):

(4) *When I get an idea I start at once building it up* [in my imagination].
And I see him [in my head].
I have to watch a baseball game [on TV].

There are fillers that are absent among the fillers of the [Locative] slot, like *imagination* or *Internet*, and there are as well classes that are present in both sets of fillers, [Locative] and [Metaphoric_Locative], like *book* or *head*. These groups can be analyzed through both DSs, although, the sense would be different: for instance, in (5a) *book* functions as a kind of informational storage, so *[in the book]* is a [Metaphoric_Locative] here, while (5b) is an example of [Locative]:

(5) *a) It is written* [in the book].
b) A dollar [in the book] *will increase the value of it*.

Unlike [Locative] fillers, classes that fill [Metaphoric_Locative] are usually used with very few prepositions: *in the imagination*, but not **under/near imagination* (of course, here we mean only their locative usage). Correspondingly, classes that fill both DSs (like *head* or *book*) also allow a full set of locative prepositions when filling the [Locative] slot and only 'default' prepositions—when filling [Metaphoric_Locative] ('default' vs 'semantic' prepositions are discussed in section 3.3 below).

As in the case with the [Locative] slot, it is convenient to mark the fillers of [Metaphoric_Locative] with a special distributional semanteme—<<MetaphoricPlace>>.

[Metaphoric_InitialPoint] and [Metaphoric_FinalPoint] correspond to [Metaphoric_Locative] in their filling, and to [Locative_InitialPoint] and [Locative_FinalPoint] in their semantics, so the former can be illustrated with examples like *Many*

details had obviously gone [from my memory], and the latter with an example like *Strange ideas come [into his mind]*.

The [Locative_Event] slot includes metonymy cases in which some event functions as a place where it occurs, for instance:

(6) *Imagine sitting* [in a meeting] *or a coffee shop.*
The artist … spent 9 years of his life [in war and imprisonment].

The fillers of the slot are classes denoting different events—*exhibition*, *rehearsal*, *parade*, *lesson*, *conference* and so on. All of them are marked with the <<Event-Place>> semanteme.

[LocativeEvent_InitialPoint] and [LocativeEvent_FinalPoint], correspondingly, have the same filling, and possess the semantics of initial and final point (as in *come [from/to the exhibition]*).

It seems reasonable to introduce three groups of locative fillers in the model for several reasons.

First, different groups of locative fillers have different syntactic realizations, namely, the fillers of [Locative], [Locative_InitialPoint] and [Locative_FinalPoint] can combine with many prepositions, like *in, at, on, under, above, near* and so on, while the fillers of metaphoric and event locatives have significant restrictions here, as shown above.

Second, different cores combine with different locatives in various ways. For example, [Metaphoric_Locative] is more typical for the verb *to read* than the [Locative] slot is, as *read [in the book/on the Internet]* is more frequent than *read [at home]*.

The following section suggests the syntactic part of the description and considers the problems bound with the syntactic pattern of the formalism.

3. The Syntactic Description of Locative Dependencies in the ABBYY Compreno Model

One of the basic elements in our syntactic module is a system of SurfSs—the syntactic positions of the constituents. For instance, the labels $Subject, $Object_Direct and $Verb in Figure 2 are SurfSs for the constituents of example (7):

(7) *The boy lives in Kiev.*

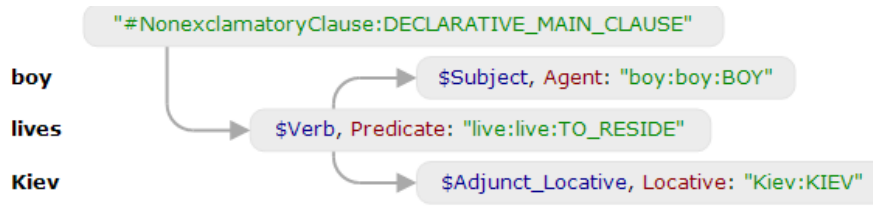


Fig. 2. Semantic and syntactic structure for example (7)

The main features of SurfSs are:

- *Government*, or the grammeme restrictions that a constituent must satisfy to be analyzed in the SurfS. For example, we can specify case forms and prepositions through grammemes, e.g., we use the grammeme of accusative case in the government of the Russian $Object_Direct SurfS.
- *Linear order* that describes the linear positions where the SurfS is allowed. For instance, the linear order for the $Object_Direct slot in English does not allow the leftmost position in a sentence, while for the $Subject slot this position is the most common.
- *Punctuation* that describes the punctuators (e.g., comma, bracket, semicolon, etc.) which are allowed in the SurfS.

If some syntactic positions share all these features, they will be reduced to one SurfS only. This reduction contributes to the productivity of the model: on the very first stage of parsing, the system builds all possible syntactic links between all words in a sentence. Each syntactic link with a child node is marked by the SurfS of the child node, so if we have fewer SurfSs, there will be fewer links generated in the analysis stage, which reduces the time necessary for the analysis.

However, the reduction of similar syntactic positions to one SurfS is not always possible. For example, when the positions have the same government but allow different linear orders such a reduction is problematic, and locative adjuncts are just the case.

3.1. Locative SurfSs. Description of the Problem

The absolute majority of Russian and English locative prepositions can be used for marking non-locative relations as well. For instance:

Russian

(8) *Таракан живет за печкой.*
Cockroach lives ZA oven.
The cockroach lives behind the oven.

(9) *Мальчик пошел за доктором.*
Boy went ZA doctor
The boy went for the doctor.

English

(10) *The book is **on** the shelf.*

(11) *Please, advise me **on** the medical training.*

As demonstrated in examples (8)–(11), the Russian preposition *za* can mark not only location but purpose as well, and the English preposition *on* can mark not only location but also theme. However, there are some syntactic differences between the locative and non-locative usage of these prepositions.

First, locatives are freely admitted in the leftmost position and this usage does not seem emphatic, as in example (12), for instance. For most non-locative groups, this position is emphatic, if admitted at all, as in (13):

(12) [In Hertford, Hereford, and Hampshire]*, **hurricanes hardly ever** happen.*

(13) *[In prepositions] *this problem is.*

Second, locatives may use other relative pronouns than non-locatives, namely, non-locative groups never use where-relativizers, whereas for locative groups this is quite normal. Compare examples (14) and (15):

(14) *The problem you advised me on. vs *The problem where you advised me.*

(15) *The city where I live.*

For these reasons we decided to introduce specific SurfSs for locative adjuncts.

3.2. Locative SurfSs and Their Government

We use the following locative SurfSs for each language: $Adjunct_Locative, $Adjunct_FinalPoint, $Adjunct_InitialPoint.

Each slot must allow nouns with specific prepositions (e.g. *to the table* for $Adjunct_FinalPoint), locative adverbs (like *below*), and adverbial pronouns (like *here* or *whence*). These instances are heterogeneous and it is not easy to cover them all in a simple description. To solve this problem we added a new grammatical category named FormOfLocativeCircumstance. It is defined for all the hierarchy, or to be more precise, for all classes that can fill locative slots, and includes the values indicated in Figure 3:

Grammemes allowed in $AdjunctLocative.
-DefaultLocativeLikeForm.
-SemanticLocativeLikeForm.
Grammemes allowed in $AdjunctlnirialPoint.
-DefaultFromForm.
-SemanticFromForm
Grammemes allowed in SAdjunctFinalPoint:
-DefaultToForm.
-SemanticToForm.

Fig. 3. Grammatical category FormOfLocativeCircumstance

The category works as follows: to indicate that a word *shelf*, for instance, can fill the $Adjunct_Locative slot with a preposition *on*, one has to indicate the preposition *on* in the DefaultLocativeLikeForm pattern of the class “shelf:SHELF”.

3.3. ‘Default’ and ‘Semantic’ Prepositions

It is possible to describe all the variety of locative adjuncts using one single pattern for each SurfS. Nevertheless, it turned out to be more convenient to split the pattern into two: ‘default’ and ‘semantic’. The ‘semantic’ pattern covers prepositions like *behind* or *near*: the semantic interpretation of their locative usage is not determined by the noun they modify, and such prepositions have exact counterparts in other languages, which can serve as translation analogues in all of the contexts (like Eng. *near*—Rus. *около*). Usually these prepositions denote peripheral spatial localizations such as AD or APUD (using the terms described in [Plungian 2000]).

The ‘default’ prepositions are quite different. They correspond to IN-Localization in the terms defined in [Plungian 2000] and form collocations with the nouns they modify. Different nouns combine with different prepositions and this choice is highly language-specific. For example, English uses different prepositions for *in the country*, *on the island* and *at the pole* while the localization denoted by the prepositions is the same.

It seems reasonable to treat the ‘semantic’ and ‘default’ prepositions differently. The ‘semantic’ prepositions can be described by one single pattern that is introduced high in the hierarchy and is very rarely modified below. This is not the case for the ‘default’ patterns—these must be described for specific lexical units.

In addition, the ‘default’ prepositions, unlike the ‘semantic’ ones, can correspond to the DSs such as [Metaphoric_Locative], as these positions denote not a real space, but a virtual one, where the localizations like APUD are simply irrelevant. One cannot talk of something near or behind the memory. This feature is not difficult to describe using two locative patterns, whereas it would be a problem for a one-pattern approach.

3.4. ‘Default’ and ‘Semantic’ Dichotomy Used for Machine Translation

The two pattern system turned out to be useful for the correct translation of locative prepositions as well. The sense of ‘semantic’ prepositions is retained through the system of transfer rules (which are discussed in [Anisimovich et al. 2012, Bogdanov Leontyev 2013]), where each preposition corresponds to a special semanteme (e.g., the preposition *under* corresponds to the semanteme <<Under>>), which demands the necessary preposition at the synthesis stage. In the case of the ‘default’ prepositions, the preposition itself is ignored, it is only the grammeme DefaultLocativeLikeForm that serves as an input for the rules. The special semanteme <<Default_Location>> is computed in this case. Using this approach we can get the correct Russian translations for the English sentences (16) and (17):

(16)	The	boy	lives	in	Kiev.	
		Мальчик	живет	в	Киеве.	

(17)	The	boy	lives	in	the	East.
		Мальчик	живет	на		Востоке.

In the English sentences (16) and (17) the single preposition *in* is used, while the Russian translations demand different prepositions—*в* and *на*. In these examples the preposition *in* is 'default', so it corresponds to the semanteme <<DefaultLocation>>. At the stage of building the output Russian structure, this semanteme will not evoke any concrete preposition but rather a link to the 'default' locative pattern—DefaultLocativeLikeForm. This pattern will assign the preposition *в* for the noun *Киев* and the preposition *на* for the noun *Восток*. Thus the resulting preposition depends only on this pattern and does not correspond directly to any preposition in the input structure.

Example (18) demonstrates the translation of 'semantic' prepositions:

(18)	The	cat	was	under	the	table.
		Кошка	была	под		столом.

Here the preposition *under* is 'semantic'. Its sense is rendered by a special semanteme <<Under>, which evokes the corresponding preposition at the synthesis stage directly, without reference to the locative patterns.

4. Conclusion

The semantic pattern of the model defines the necessary DSs for the locative valencies, sets their filling and introduces the slots in the hierarchy. The syntactic pattern provides the description of their syntactic realization, differentiating between the 'semantic' prepositions and the 'default' ones. The necessity of introducing specific SurfSs for the locative adjuncts as well as additional restrictions for them makes the work of the parser more complicated, but we believe that these difficulties reflect an objective complexity of language. Besides, such a model allows one to achieve a full and integral description of locative dependencies, as has already been done for the English and Russian locative constructions. Now the system is to be tested on a wider range of languages.

References

1. *Anisimovich, K. V., K. Y. Druzhkin, F. R. Minlos, M. A. Petrova, V. P. Selegey and K. A. Zuev.* (2012) Syntactic and Semantic Parser Based on ABBYY Compreno Linguistic Technologies, Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2012). [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2012] Bekasovo, 91–103.
2. *Aurnague, Michel Hickmann, Maya & Vieu, Laure.* (eds.) (2007) The categorization of spatial entities in language and cognition, Amsterdam, John Benjamins.
3. *Bogdanov, A. V., Leontyev A. P.* (2013) Description of the Russian External Possessor Construction in a Natural Language Processing System, Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2013). [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2013] Bekasovo, 110–118.
4. *Bloom P., Peterson M. A., Nadel L., Garrett, Merrill F.* (1996) Language and space, Cambridge, Mass., MIT Press.
5. *Feigenbaum S., Kurzon D.* (eds.) (2002) Prepositions in their Syntactic, Semantic and Pragmatic Context. [= Typological studies in language 50]. Amsterdam, Philadelphia: John Benjamins.
6. *Filipenko M. V.* (2000) The problems of description of prepositions in the modern linguistic theories [Problemy opisaniya predlogov v sovremennykh lingvisticheskikh teoriyakh]. In Paillard D. Seliverstova O. N. (eds) (2000) Research on the semantics of prepositions [Issledovaniya po semantike predlogov]. Moscow, Russian Dictionaries.
7. *Gak V. G.* (1996) The functional semantic field of the predicates of localization [Funkcional'no-semanticheskoye pole predicatov lokalizacii], The theory of functional grammar. Locativity. Existentiality. Possessivity. Conditionality [Teoriya funkcional'noy grammatiki. Lokativnost'. Bytiynost'. Posessivnost'. Obuslovlennost'], St-Petersburg, 6–26.
8. *Gak V. G.* (2000) Space outside space [Prostranstvo vne prostranstva]. Logical analysis of the language [Logicheskiy analis yazyka], Moscow Languages of Russian culture 127–134.
9. *Hickmann M., Robert S.* (eds.) (2006) Space in Languages: Linguistic Systems and Cognitive Categories. Amsterdam / Philadelphia: John Benjamins.
10. *Levinson, Stephen C. & Wilkins David P.* (ed.) (2006) Grammars of space, Cambridge, Cambridge UniversityPress.
11. *Levinson, Stephen C.* (2003) Space in language and cognition : explorations in cognitive diversity, Cambridge, Cambridge University Press.
12. *Manicheva, E., M. Petrova, E. Kozlova and T. Popova.* (2012) 'The Compreno Semantic Model as Integral Framework for Multilingual Lexical Database.' In Zock, M. and R. Rapp (eds), Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III), COLING 2012. Mumbai, 215–229.
13. *Petrova M. A. (2013) The Compreno Semantic Model: The Universality Problem //* International Journal of Lexicography 2013; doi: 10.1093/ijl/ect038 (in print)

14. *Plungian V. A.* (2000) General morphology. Introduction to the problems [Obschaya morphologiya. Vvedeniye v problematiku]. Moscow, Editorial URSS.
15. *Shay, Erin & Seibert, Uwe* (2003) Motion, direction and location in languages, Amsterdam—Philadelphia, John Benjamins.
16. *Svorou S.* (1994) The grammar of space, Amsterdam—Philadelphia, John Benjamins.
17. *Shvedova N. Yu.* (ed) (1980) Russian grammar, Moscow.
18. *Talmy L.* (1983) How language structures space. In Herbert L. Pick, Jr. & Linda P. Acredolo (eds.), Spatial orientation: Theory, research, and application, 225–282. New York: Plenum Press
19. *Vsevolodova M. V.* (2010) Grammatical aspects of the Russian prepositional entities: typology, structure syntagmatics and syntactic modifications [Grammaticheskiye aspekty russkikh predlozhnykh yedinic: tipologiya, struktura, sintagmatika i sinaksicheskiye modifikacii] In Problems of Linguistics [Voprosy yazykoznaniya] vol 4. pp. 3–26.
20. *Vsevolodova M. V., Vladimirskiy Ye. Yu.* (1982) Means of expression of the spatial relations in the modern Russian language [Sposoby vyrazhaniya prostranstvennykh otnosheniy v sovremennom russkom yazyke], Moskow
21. *Van der Zee, Emile & Slack, Jon,* eds (2003) Representing direction in language and space, New York, Oxford University Press.
22. *Zuyev, K. A., E. M. Indenbom and M. V. Yudina,* (2013) 'Statistical Machine Translation with Linguistic Language Model.' Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferencii 'Dialog' 2013. Bekasovo, 164–172.

УНИВЕРСАЛЬНЫЕ МЕЛОДИЧЕСКИЕ ПОРТРЕТЫ ИНТОНАЦИОННЫХ КОНСТРУКЦИЙ РУССКОЙ РЕЧИ

Лобанов Б. М. (Lobanov@newman.bas-net.by),
Окрут Т. И. (tatberrie@gmail.com)

Объединённый институт проблем информатики
НАН Беларуси, Минск, Беларусь

Ключевые слова: интонационные конструкции, мелодический портрет, синтез и анализ интонации, русская интонации, русский как иностранный (РКИ)

UNIVERSAL MELODIC PORTRAITS OF INTONATION PATTERNS IN RUSSIAN SPEECH

Lobanov B. M. (Lobanov@newman.bas-net.by),
Okrut T. I. (tatberrie@gmail.com)

United Institute of Informatics Problems NAS Minsr, Belarus

We proceed from the model of intonation patterns by Elena Bryzgunova, which is widely used in the teaching of Russian speech intonation. This model includes seven patterns: IP1 (the falling tone), IP2 (the falling tone with a certain prosodic emphasis), IP3 (the rising tone with subsequent fall), IP4 (the falling-rising tone). IP5 (combination of the rising, smooth and falling tones), IP6 (combination of the rising and smooth tones), IP7 (combination of the rising tone with the glottal stop). We present a model of intonation portraits of accentual units (the PAU model), proposed by one of the authors of this paper and effectively used in the practice of Russian speech synthesis for a long time. The PAU model assumes that, for a certain intonation type, the topological properties of the melodic contour are independent of the quantitative and the qualitative characteristics of the pre-nucleus, the nucleus and the post-nucleus of accentual units. The methodology of an experiment of integration of the two models into a unified model of Universal melodic portraits of intonation patterns (UMP-IP) is discussed. The new model is shown to effectively represent the tonal structure of Elena Bryzgunova's intonation patterns and ensure the invariance of the quantitative and the qualitative constituents of the sentence pronounced as well as the pitch and the range of the speaker's voice. The obtained results are discussed from the viewpoint of applicability to the practice of teaching Russian as the second language.

Key words: intonation patterns, melodic portrait, synthesis and analysis of intonation, Russian intonation, Russian as the second language

Введение

В 1960-х гг. Е. А. Брызгунова предложила описание интонации русского языка [Брызгунова, 1968] с использованием понятия *интонационной конструкции*, которое вошло в академическое издание русской грамматики [Брызгунова, 1980] и стало повсеместно использоваться методических пособиях по обучению русского языка как иностранного (РКИ) [Одинцова, 2011]. За основу классификации интонационных конструкций принят характер движения тона на ударном и прилегающих к нему слогах. Е. А. Брызгунова выделяет семь основных интонационных конструкций русского языка, различающих смысл звучащих предложений. Приведём особенности их типичного употребления и описание характера движения тона в каждой ИК, взятые из современного Интернет-учебника по фонетике русского языка [http://www.philol.msu.ru/~fonetica/]:

- ИК1 наблюдается при выражении завершенности в повествовательных предложениях: *Анна стоит на мосту. Наташа поет*. Для ИК1 характерно понижение тона на ударной части.
- ИК2 реализуется в вопросе с вопросительными словами: *Кто пьет сок? Как поет Наташа?* При ИК2 ударная часть произносится с некоторым повышением тона.
- ИК3 характерна для вопроса без вопросительного слова: *Это Антон? Ее зовут Наташа?* Для этой интонации характерно значительное повышение тона на ударной части.
- ИК4 — это вопросительная интонация, но с сопоставительным союзом /а/: *А вы? А это?* На ударной части происходит повышение тона, продолжающееся на безударных слогах.
- ИК5 реализуется при выражении оценки в предложениях с местоименными словами: *Какой сегодня день!* На ударной части — повышение тона.
- ИК6, так же как и ИК5, реализуется при выражении оценки в предложениях с местоименными словами: *Какой сок вкусный!* Повышение тона происходит на ударной части и продолжается на заударной части.
- ИК7 употребляется в предложениях со значением экспрессивного отрицания признака, действия, состояния: *Какой он специалист! Только вид делает*.

Очевидно, что, представленное выше описание ИК1-ИК7, также как и более развёрнутое описание приведенное в [Брызгунова, 1980], не являются полными и строгими. Эти описания понятны лингвистам и преподавателям РКИ, но не могут вполне удовлетворить разработчиков компьютерных моделей анализа и синтеза интонационных характеристик речи.

В настоящей работе предпринята попытка дать описание ИК1-ИК7 в рамках хорошо разработанной и апробированной ПАЕ-модели [Lobanov, 2006]. ПАЕ-модель (модель портретов акцентных единиц) была предложена более 20 лет назад [Lobanov, 1987] и с тех пор успешно использовалась в нескольких системах синтеза русской речи по тексту. В соответствии с ПАЕ-моделью, минимальной просодическим компонентом, из которого составляется интонация синтагмы, является Акцентная Единица (АЕ). АЕ может состоять из одного или

более фонетических слов, но должна иметь в своём составе только один полноударный слог. Каждая АЕ, в свою очередь, состоит из ядра (полноударная гласная фонема), предъядра (все фонемы, предшествующие полноударной гласной) и заядра (все фонемы за полноударной гласной). Или иначе — согласно терминологии Брызгуновой — «центр», «предцентр» и «постцентр». ПАЕ-модель предполагает, что для определенного типа интонации топологические свойства мелодического контура АЕ не зависят от количественного и качественного содержания предъядра, ядра и заядра.

ПАЕ-модель обеспечивает возможность представления семи интонационных конструкций Брызгуновой — {ИК$_i$} — в виде набора их Универсальных Мелодических Портретов (ЭМП) в нормированных координатах «Частота — Время» — {УМП ИК$_i$}.

Нормализация по времени осуществляется путём приведения к стандартной длине элементов АЕ: предъядерных, ядерных и заядерных участков. Этот вид нормализации устраняет различия мелодической кривой, связанные с количественным составом предядерных и заядерных участков АЕ.

Для нормализации по частоте определяются минимальное ($F_{0\,min}$) и максимальное ($F_{0\,max}$) значения частоты основного тона — F_0 для всего ансамбля мелодических кривых {ИК$_i$} в произнесении данного диктора. Нормализация осуществляется в соответствии с формулой: $F_0^N = (F_0 - F_{0\,min}) / (F_{0\,max} - F_{0\,min})$. Этот вид нормализации устраняет различия мелодической кривой, связанные с индивидуальными дикторскими различиями в высоте голоса.

Таким образом, нормированное пространство для отображения УМП ИК$_i$ может быть представлено в виде прямоугольника с координатными осями (T_N, F_0^N), как это представлено на схематическом рисунке 1. При этом интервалам на оси абсцисс соответствуют: [0–1/3] — предъядро, [1/3–2/3] — ядро, [2/3–1] — заядро. Интервалам на оси ординат соответствуют: [0–1/3] — низкий уровень тона, [1/3–2/3] — средний, [2/3–1] — высокий.

Целью настоящей работы является экспериментальная проверка эффективности представления ИК1 — ИК7 в виде набора Универсальных Мелодических Портретов — {УМП ИК$_i$} — в условиях варьирования количественного и качественного состава произносимых фраз, а также в условиях произнесения этих фраз женским и мужским голосами.

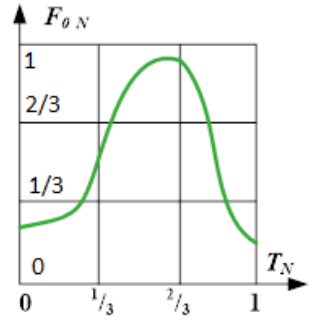


Рис. 1. Общий вид УМП ИК

1. Методика эксперимента

В качестве материала для экспериментального исследования использованы аудиозаписи многочисленных вариантов эталонной реализации каждой из семи ИК в исполнении профессиональных дикторов (женщина и мужчина). Аудиозаписи взяты из приложения к учебному пособию РКИ [Одинцова, 2011].

С помощью системы ***PhonoClonator*** [Лобанов, 2008] для каждого из аудио сигналов эталонной реализации ИК осуществлялась автоматическая сегментация и разметка на последовательность аллофонов фонем и питчей (периодов основного тона). Разметка аудио сигнала осуществлялась в соответствии с текстом анализируемой фразы. В таблице 1 приведены тексты фраз в каноническом виде (как в учебнике) и в виде, требуемом для работы ФоноКлонатора. Общий вид пользовательского интерфейса ФоноКлонатора при обработке этих фраз приведен на рисунке 2.

Таблица 1

	Канонический текст фразы с ИК1	**Текст фразы с ИК1 для ФоноКлонатора**
1.	*Он гул[1]яет.*	*О=н гуля+ет.*
2.	*Он гул[1]яет в парке.*	*О=н гуля+ет в па=рке.*
3.	*Он гул[1]яет в парке с собакой.*	*О=н гуля+ет в па=рке с соба=кой.*
4.	*Он заним[1]ается.*	*О=н занима+ется.*
5.	*Он заним[1]ается в библиотеке.*	*О=н занима+ется в библиоте=ке.*

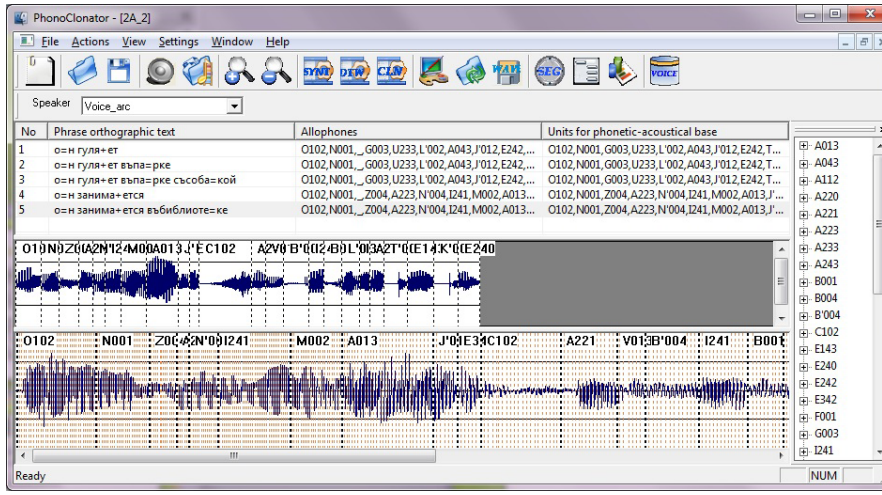


Рис. 2. PhonoClonator — общий вид пользовательского интерфейса

На следующем этапе размеченные ФоноКлонатором аудио сигналы подаются на вход системы ***IntoClonator*** [Лобанов, 2008], с помощью которого отображаются

границы ядра, предъядра и заядра, мелодическая кривая и кривая интенсивности сигнала (см. рисунок 3). Предварительно устанавливаются минимальное ($F_{0\,min}$) и максимальное ($F_{0\,max}$) значения частоты основного тона —F_0 для всего ансамбля мелодических кривых {ИК$_i$} в произнесении данного диктора. В приведенном на рис. 3 примере для фразы с ИК1 «*Он занима́ется в библиотеке»,* произнесённой диктором женщиной выбраны: $F_{0\,min}$ = 115 Гц, а $F_{0\,max}$ = 420 Гц.

Рис. 3. IntoClonator — общий вид пользовательского интерфейса

На заключительном этапе с помощью системы ***ShapeEditor*** (см. Рис. 4) обработанная *ИнтоКлонатором* информация используется для построения УМП ИК1 для фразы «*Он занима́ется в библиотеке*» в описанном выше нормированном виде.

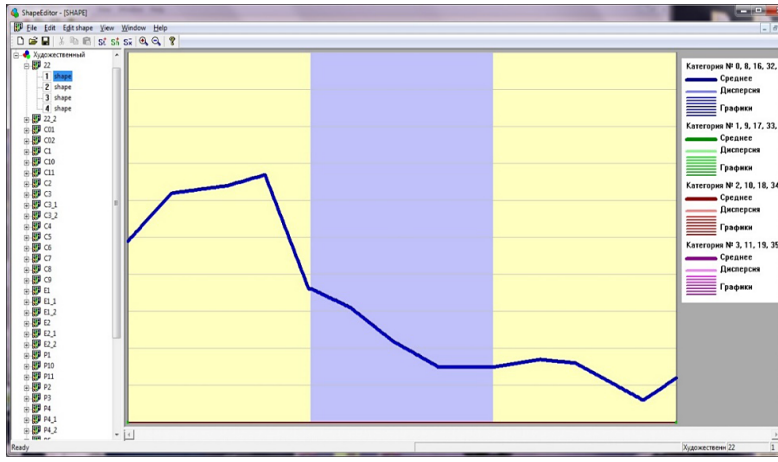
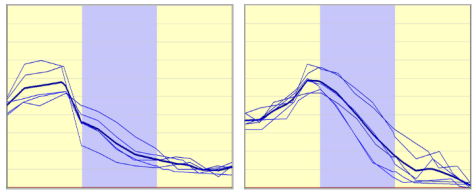


Рис. 4. ShapeEditor — общий вид пользовательского интерфейса

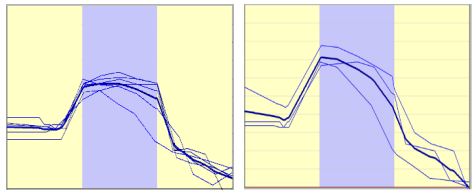
2. Результаты эксперимента

Ниже на рисунках 5–11 приведены построенные в соответствии с описанной выше методикой УМП ИК1 — ИК7 для различных фраз, произнесённых диктором-мужчиной (слава) и диктором-женщиной (в центре). Справа приведена таблица с текстами анализируемых фраз. Для мужского и женского голосов при нормировании частотной шкалы для всех ИК использовались, соответственно, значения $F_{0\,min}$ = 75 и 115 Гц, определённые для контуров ИК1, и значения $F_{0\,max}$ = 210 и 420 Гц, определённые для контуров ИК5. На каждом из построенных графиков УМП жирная кривая описывает усреднённые данные для всех использованных реализаций ИК.



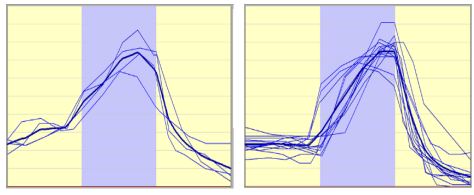
1	Он гул[1]яет.
2	Он гул[1]яет в парке.
3	Он гул[1]яет в парке с собакой.
4	Он заним[1]ается.
5	Он заним[1]ается в библиотеке.

Рис. 5. ИК1 — УМП мужского и женского голосов + тексты фраз



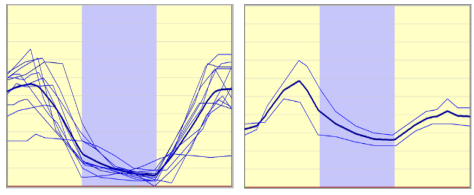
1	Как[2]ой?
2	Как[2]ой фильм?
3	Как[2]ой вчера был фильм?
4	Как[2]ой вчера был фильм по телевизору?

Рис. 6. ИК2 — УМП мужского и женского голосов + тексты фраз



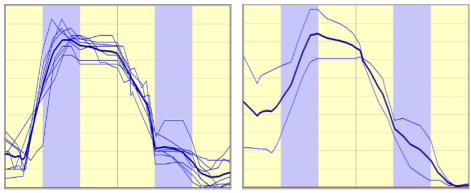
1	Ты своб[3]оден сегодня?
2	Ты свободен сег[3]одня?
3	Л[3]юбит?
4	Сестра л[3]юбит?
5	Сестра л[3]юбит балет?

Рис. 7. ИК3 — УМП мужского и женского голосов + тексты фраз



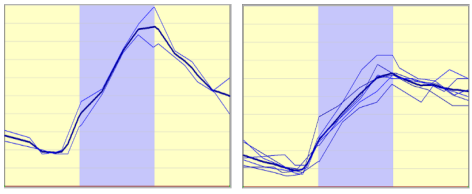
1	А д[4]едушка?
2	А ч[4]ёрный?
3	А в м[4]алом?
4	А р[4]озы?
5	А сег[4]одня?

Рис. 8. ИК4 — УМП мужского и женского голосов + тексты фраз



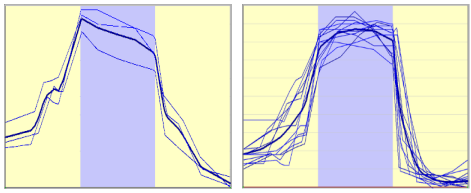
1	Как[5]ой сегодня д[5]ень!
2	Как[5]ой сегодня ужасный д[5]ень!
3	Гд[5]е он только не побыв[5]ал!
4	Ком[5]у она только не пис[5]ала!

Рис. 9. ИК5 — УМП мужского и женского голосов + тексты фраз



1	Какой ф[6]ильм!
2	Какой ф[6]ильм прекрасный!
3	Какая м[6]узыка прекрасная!
4	Какая у них м[6]ебель!
5	Какая у них м[6]ебель некрасивая!

Рис. 10. ИК6 — УМП мужского и женского голосов + тексты фраз



1	Да как[7]ая там выставка!
2	Да как[7]ая она актриса!
3	Да оъч[7]ём мне с тобой разговаривать!
...	...

Рис. 11. ИК7 — УМП мужского и женского голосов + тексты фраз

Представленный на рисунках 5–11 экспериментальный материал позволяет сделать главный основополагающий вывод о том, что УМП являются эффективным средством представления тональной структуры интонационных конструкций Е. А. Брызгуновой, обеспечивающим инвариантность относительно количественного и качественного состава произносимых фраз, а также высоты и диапазона голоса диктора.

На рисунке 12 представлены УМП ИК1 в сопоставлении с ИК2 — ИК7. Очевидность графического сопоставления не требует, на наш взгляд дальнейших словесных пояснений. Суждение о том, в какой степени полученные УМП ИК1–7 коррелируют с данными, приведенными в [Брызгунова, 1980] (см. таблицу), мы предоставляем читателю.

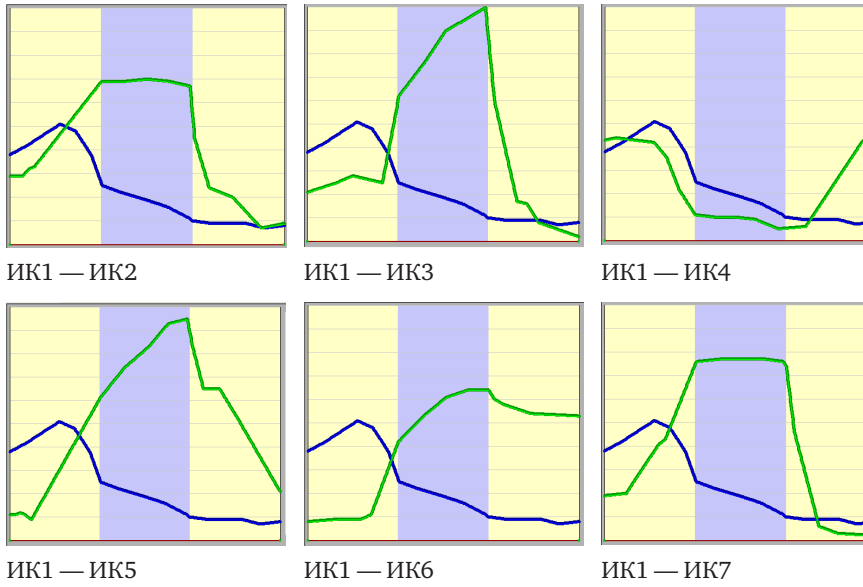


Рис. 12. Графическое сопоставление УМП ИК1 (тёмная линия) с ИК2 — ИК7

Таблица 2

Тип ИК		Направление тона в центре ИК	Уровень тона центре	Уровень тона в постцентре	Признаки для 2-х типов ИК
¯¯_	ИК-1	Нисходящее	Ниже предцентра	Ниже предцентра	См. ИК-2
––_	ИК-2	Нисходящее	В пределах предцентра или незначительно ниже	Ниже предцентра	Усиление словесного ударения на гласном центра в отличие от ИК-1
––/_	ИК-3	Восходящее	Выше предцентра	Ниже предцентра	См. ИК-7
__/¯	ИК-4	Нисходящее или нисходяще-восходящее	Ниже предцентра	Выше предцентра	
–/¯_	ИК-5	1-й центр — восходящее, 2-й центр — нисходящее	Выше предцентра	Ниже предцентра	Увеличение длительности центров по сравнению с ИК-2
__/¯	ИК-б	Восходящее	Выше предцентра	Выше предцентра	
_ _/_	ИК-7	Восходящее	Выше предцентра	Ниже предцентра	Смычка голосовых связок на гласном центра в отличие от ИК-3

3. Обсуждение результатов

В данной работе мы сознательно ушли от рассмотрения вопросов, связанных с употреблением различных ИК для выражения смысловых различий высказываний во взаимодействии с их синтаксическим строением и лексическим составом, сосредоточив внимание лишь на представлении акустических компонентов интонации. Строгое решение вопросов ИК-разметки текстов до сих пор остаётся открытым, хотя постоянно находятся в сфере научных интересов [Янко, 2008]. Тем не менее, уже существует большое количество текстов в различных учебниках РКИ, размеченных «вручную» с указанием требуемых ИК. Ниже приведен пример такого рода текста из учебного пособия [Муханов, 1995]. В приведенном тексте центр ИК маркируется цифрой, соответствующей типу ИК. Ядру УМП соответствует маркированная цифрой ударная гласная фонема, а левой (предъядро) и правой (заядро) границами УМП являются знаки препинания либо знак {/}.

Наташа: Здра^{2}вствуй, Петя! Ну, удало^{3}сь тебе купить книгу?
Петя: Здра^{2}вствуй, Наташа! Да^{1},/ я купи^{1}л книгу. Хорошо, что ты посоветовала в Дом кни^{1}ги. Зна^{3}ешь, что я ещё там купил?
Н.: Что^{2}?
П.: «Антоло́гию русской поэ^{2}зии»!
Н.: Пра^{3}вда? Во^{3}т здорово! Ты не мо^{3}г бы дать мне её почитать?
П.: Почему^{2} же не могу? Обяза́тельно да^{1}м,/ но только попо^{1}зже.
Н.: Ну, хорошо^{1}. А больше ты ничего^{3} не купил?
П.: Я хотел купить ещё альбом «Пейзажи Москвы^{1}»,/ но у меня не хвати́ло де^{1}нег.
Н.: Так ведь у меня е^{2}сть этот альбо́м. Хо^{3}чешь, дам посмотреть?
П.: Конечно, хочу^{2}. Когда^{2} ты сможешь мне его дать?
Н.: Да хоть сейча^{2}с,/ он у меня с собо^{2}й.
П.: Ну, прекра^{1}сно. Спаси^{2}бо тебе большое. Ну, до за^{3}втра?
Н.: До за^{2}втра.

Отметим важное практическое значение разработанного метода представления интонации речи в виде набора УМП ИК при создании компьютерных программ обучения русскому языку как иностранному с использованием синтезаторов и анализаторов речи. УМП ИК могут быть использованы в составе русскоязычных синтезаторов речи для освоения русской интонации на слух и в составе анализаторов речи — для тренировки правильности интонирования в процессе говорения.

Если в соответствии с разработанной методикой (см. раздел 1) создан эталонный набор УМП ИК и имеется текст, размеченный приведенным выше способом, то он может быть озвучен с желаемыми интонационными характеристиками. Для этой цели может быть использован синтезатор речи *MultiPhone* [Лобанов, 2008], в котором созданные УМП ИК применяются в качестве эталонов. При включении его в состав обучающих программ РКИ преподаватель или учащийся сможет самостоятельно формировать тренировочные упражнения

для освоения звучания различных ИК русской речи, в различных контекстах и с различными голосами.

Созданный эталонный набор УМП ИК может быть использован также в анализаторе интонации произносимых фраз. Для этой цели создаётся специализированный программный комплекс, последовательно реализующий описанные в разделе 1 функции 3-х систем: *PhonoClonator* + *IntoClonator* + *ShapeEditor* [Лобанов, 2008]. С помощью этого программного комплекса произносимая учащимся в микрофон фраза анализируется и на основе анализа спектральных, мелодических и энергетических характеристик строится её УМП. Затем полученный УМП сравнивается с эталонным и по результатам сравнения даётся оценка правильности интонирования произнесённой фразы, а также рекомендации по исправлению интонирования предъядра, ядра или заядра фразы.

Предложенный метод представления мелодического контура в виде УМП может найти также эффективное применение в различного рода сопоставительных лингвистических исследований интонации. Метод УМП позволяет получить более строгие оценки сходства или различия интонационных феноменов индивидуальности дикторского чтения, особенностей интонационных стратегий в исследуемых языках и их диалектах.

Литература

1. *Брызгунова Е. А.* (1968) Звуки и интонация русской речи / — Наука, М.
2. *Брызгунова Е. А.* (1980) Интонация // Русская грамматика / — Наука, М.
3. *Lobanov B.* (1987) The phonemophon text-to-speech system // International Congress of Phonetic Sciences: proc. of the 11th seccion ICPhS'87, — Tallin, — V. 1. — P. 120–124.
4. *Lobanov B.* (2006) Language- and Speaker Specific Implementation of Intonation Contours in Multilingual TTS Synthesis / Lobanov B., Tsirulnik L., Zhadinets D., Karnevskaya E. // Speech Prosody: proceedings of the 3rd International conference. Dresden, Germany — Vol. 2. — P. 553–556.
5. *Лобанов Б. М.* (2008) Компьютерный синтез и клонирование речи / Б. М. Лобанов, Л. И. Цирульник // Белорусская Наука, Минск. (См. также: http://www.pselab.ru/Books/Lobanov_Cirulnik_2008.pdf)
6. *Муханов И. Л.* (1995) Интонация в практике русской диалогической речи / Ойкумена, М.
7. *Одинцова И. В.* (2011) Звуки. Ритмика. Интонация. М.: Флинта-Наука, 2011.
8. *Янко Т. Е.* (2008) Интонационные стратегии русской речи в сопоставительном аспекте / Языки славянских культур, М.

РУТЕЗ-LITE, ОПУБЛИКОВАННАЯ ВЕРСИЯ ТЕЗАУРУСА РУССКОГО ЯЗЫКА РУТЕЗ

Лукашевич Н. В. (louk_nat@mail.ru),
Добров Б. В. (dobrov_bv@mail.ru),
Четверкин И. И. (ilia2010@yandex.ru)

Московский государственный университет
им. М. В. Ломоносова, Москва, Россия

Ключевые слова: автоматическая обработка текстов, WordNet, тезаурус, толкования

RUTHES-LITE, A PUBLICLY AVAILABLE VERSION OF THESAURUS OF RUSSIAN LANGUAGE RUTHES

Loukachevitch N. V. (louk_nat@mail.ru),
Dobrov B. V. (dobrov_bv@mail.ru),
Chetviorkin I. I. (ilia2010@yandex.ru)

Lomonosov Moscow State University, Moscow, Russia

The paper presents RuThes-lite, a publicly available version of RuThes linguistic ontology, which has been developed for more than fifteen years and is intended for automatic document processing. RuThes has considerable similarities with WordNet: 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 names of concepts, attention to multiword expressions, intentional inclusion of terms of the sociopolitical domain, a set of conceptual relations. RuThes-lite was generated from RuThes on the basis of the most frequent words in a contemporary news collection. Besides, we describe additional data, which have been specially prepared for RuThes-lite publication: morph-syntactic labeling of thesaurus text entries and assignment of glosses to concepts.

Keywords: natural language processing, WordNet, thesaurus, part-of-speech tagging, gloss

Introduction

WordNet is one of popular resources used for natural language processing and information-retrieval applications (Fellbaum, 1998). For many languages projects on developing national wordnets have been initiated. At least four attempts to create a Russian wordnet are known (Azarowa, 2008; Gelfenbeyn et al., 2003; Balkova et al., 2008; Braslavski et al. 2013).

In spite of its popularity in computational linguistics applications, WordNet initially was created as a justification of a psycholinguistic theory (Miller, 1998), its structure and relations were based on psycholinguistic experiments and were not initially intended for natural language processing tasks. So, some constructive features of WordNet hinder its applications in automatic text processing.

These problems include: the initial absence of relations between different parts of speech with the same meaning (*adopt—adoption*—now this problem is corrected with special relations (Clark et al., 2008)); the absence of links between semantically related senses of derivate words (*initiation—initiator*); so-called "tennis problem", indicating the absence of relations between synsets of the same domain (*plane—airport*); problems in introducing synsets for multiword expressions. Some of these problems are partially corrected with the generation of additional data. For example, for "tennis problems"—the system of WordNet domains has been developed (Bentivogli et al., 2004; Bhatt et al., 2014), in many wordnets derivational links between synsets for labeling word derivations have been introduced (Azarova et al., 2002; Koeva et al., 2008).

Research on better structures of computer-oriented language resources is not a simple task because one should not only create a quite large resource, but also demonstrate its quality and characteristics of its structure in various NLP applications.

In this paper we will describe the structure and the current state of newly published RuThes-lite linguistic ontology, which is intended for use automatic text processing of Russian documents. RuThes-lite is a public part of RuThes ontology, which has been developed since 1994 and was applied in several tasks of natural language processing and information retrieval (Loukachevitch, Dobrov, 2014). In contrast to WordNet, in RuThes we implemented a unified representation for different parts of speech, lexical units and domain terms, single words and multiword expressions, adopted a set of conceptual relations, tested in applications.

The structure of this paper is as follows. In Section 1 we briefly describe the structure of RuThes linguistic ontology. Section 2 explains how RuThes-lite was generated from RuThes. In Section 3 we describe additional linguistic information that was specially prepared and provided for RuThes-lite. Section 4 reports some details on RuThes-lite publication.

1. RuThes Linguistic Ontology

RuThes Thesaurus of Russian language can be called a linguistic ontology for natural language processing, i.e. an ontology, where the majority of concepts are

introduced on the basis of actual language expressions. RuThes is a hierarchical network of concepts. Each concept has a name, relations with other concepts, a set of language expressions (words, phrases, terms), whose meanings correspond to the concept.

In RuThes, a unit is presented not by a set of similar words or terms, as it is done in the WordNet thesaurus, but by a concept—as a unit of thought, which can be associated with several synonymic language expressions. Every concept should have distinctions from related concepts that are independent from context and should be expressed in a specific set of relations or associated language expressions—text entries.

Each concept should have a concise and unambiguous name. Such names often help to express, delimit the denotational scope of the concept. Besides, the names facilitate the analysis of the results of natural language processing. If necessary, a concept may have a gloss, which is not a part of the concept name.

Words and phrases, which meanings are represented as references to the same concepts of the thesaurus, are called ontological synonyms. Ontological synonyms can comprise:

- words belonging to different parts of speech (*стабилизация* (*stabilization*), *стабилизировать* (*stabilize*), *стабилизационный* (*stabilizing*))—therefore the number of RuThes concepts is approximately 2.5 times less than in a wordnet-like resource of the same size;
- language expressions relating to different linguistic styles, genres;
- idioms and even free multiword expressions (for example, synonymous with single words).

A row of ontological synonyms can include quite a large number of words and phrases. So, a concept *ДУШЕВНОЕ СТРАДАНИЕ* (*wound in the soul*) has more than 20 text entries including such as: *боль*, *боль в душе*, *в душе наболело*, *душа болит*, *душа саднит*, *душевная пытка*, *душевная рана*, *душевный недуг*, *наболеть*, *рана в душе*, *рана в сердце*, *рана души*, *саднить* (several English translations may be as follows: *wound*, *emotional wound*, *pain in the soul* etc.).

Introducing a concept linguists specially search for multiple lexical variants (especially multiword ones) that can express the same sense. An introduced text entry should have the sufficient frequency in contemporary text collections. With this aim usually Yandex.news service or Yandex search engine are used. We do not use Russian National corpus for this check because it does not comprise necessary volumes of lexical data.

An ambiguous word is assigned to several concepts—this is the same approach as in WordNet. For example, word *коса* is assigned to three different concepts:

- *КОСА* (*ВЫСТУП ЗЕМЛИ*) (*tongue of land*)
- *КОСА ВОЛОС* (*braid of hair*)
- *КОСА* (*СЕЛЬСКОХОЗЯЙСТВЕННОЕ ОРУДИЕ*) (*scythe*)

Language expressions whose sense can serve as a basis for a separate concept in RuThes belong not only to the general vocabulary, but also can be terms of specific subject domains within the broad scope of social life (economy, law, international

relations, politics, transport, banks, etc.), so-called *sociopolitical domain* (Loukachevitch and Dobrov, 2004). This was done because many professional concepts, terms, and slang of these domains penetrate easily into the general language, and can be widely discussed in mass media: news reports and newspaper articles. The appearance of these terms in general news is not accidental. People interact with professionals and professional domains in everyday life and therefore should possess relevant terminology. In addition, such a scope of concepts facilitates the application of RuThes in specialized subdomains of the broad socio-political domain. Examples of such concepts in RuThes include: *EMERGENCY LOAN*, *TAX EXEMPTION*, *IMPORT TAX*, *DEMOGRAPHIC INDICATOR* etc.

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). 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 (so tree can grow in many places 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: *asc1* is directed to the main concept, *asc2*—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.

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 names of concepts, attention to multiword expressions, intentional inclusion of terms of the sociopolitical domain, the set of conceptual relations. 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.

2. Generating RuThes-Lite

We decided to publish partially RuThes creating RuThes-lite version, which includes approximately 100,000 unique text entries. Such a resource should contain the

most frequent words of contemporary Russian and at the same time include the upper levels of the RuThes hierarchy to preserve its property to be a connected net.

Frequency estimation of words is based on a news collection. Automatic news flow processing is one of the most important directions of natural language processing technologies. News and newspaper articles are categorized, clustered, from them named entities, relations, facts, opinions are extracted, special news services collect, process them and provide access to news data. In addition to news, such collections also contain newspaper articles, legal acts, and even literature pieces published in newspapers and journals.

The used news collection comprised 2 millions newspaper articles and news reports from around 2000 news sources. So RuThes text entries were matched with texts of this text collection and the revealed text entries were ordered by frequency decrease.

The beginning of the obtained list was cleaned up from compositional text expressions (usually synonymic variants of single words), names of persons and organizations, professional legal or economy terms. From this cleaned list we selected approximately 30 thousand the most frequent text entries (in fact, it was an iterative procedure), most of them were single words.

The frequency list begins quite traditionally: *быть*, *год*, *сообщать*, *мочь*, *время*, *стать*, *слово*. At the end of the list the following words are situated: *биофизика*, *абонементный*, *чаевые*, *спиваться*, *распашной* etc.

These selected text entries were used as seeds for concept extraction. In RuThes-lite the following concepts were included:

- All concepts having text entries from the seed list—seed concepts,
- Upper level concepts to seed concepts, that is concepts, which have a path of hyponymy or part relations to the selected concepts.
- For extracted concepts all their text entries and relations between each other are also extracted. The current version of RuThes-lite contains 26,365 concepts, around 96,941 unique words and expressions, 115,349 senses (concept-text entry links), 108,000 relations between concepts.

3. Preparing Additional Data for RuThes-Lite

The basic data of RuThes comprise:

- the list of thesaurus concepts including the concept identifier and its name,
- the list of text entries of thesaurus in the dictionary form and in the lemmatized form (each word in a text entry is lemmatized);
- the list of relations between text entries and concepts;
- the list of relations between concepts.

For the public version we prepared additional data useful for applications. In this paper we describe two type of additional information: morpho-syntactic labels of thesaurus text entries and glosses extracted from Wiktionary and assigned to thesaurus concepts.

Below we will describe techniques utilized for preparing these data for RuThes-lite.

3.1. Morpho-Syntactic Labeling

As indicated above, in RuThes all parts of speech, single words and multiword expressions are presented as text entries to the same concept. Each text entry is provided with the representation as a sequence of lemmas—words in dictionary forms (lemmatic representation): for example, *голубые фишки—голубой фишка*. This information was introduced manually. The part-of-speech tags of text entries were absent because it was supposed that part-of-speech labeling is produced during automatic text processing with a morphological tagger. However, for many applications information about the part of speech of a text entry, the head word of a multiword word expression can be essential.

In RuThes-lite we provide additional morphological and syntactic information about a text entry: the part of speech of a single word; the head of a phrase and the part of speech of a text entry as a whole (= part of speech of its head word) for a multiword expression.

The labeling was fulfilled automatically with morphological processing of a text entry and its lemmatic representation—the use of the both types of information decreases potential morphological ambiguity.

So, now such text entry as *уголовное дело* (*criminal case*) has the following information about own structure:

уголовное дело:
уголовный дело (words in lemmatic forms)
NG (noun group)
дело (head word)
Adj N (parts of speech for every word in the text entry).

It should be noted that word *"дело"* is morphologically ambiguous but the description eliminates the ambiguity.

3.2. Assignment of Glosses to Concepts

In RuThes names of concepts play an important role. They should be clear and unambiguous and should inform a native speaker about the meaning of a concept. Only for a small number of concepts some additional explanations are provided.

But in WordNet-like resources glosses are often used as prominent information in various applications, for example, generation of a sentiment vocabulary (Baccianella et al., 2010), calculation of similarity measures (Pedersen et al., 2004), lexical disambiguation (Agirre, Soroa 2009) and others. Therefore some wordnet developers try to mine glosses from lexical resources (Henrich et al., 2011).

For RuThes-lite we also made the first step in providing concepts with glosses explaining their intended meanings—we automatically extracted glosses from Russian Wiktionary, matched glosses and concepts and selected the most appropriate gloss for a concept. The problem here is how to select the best gloss describing the meaning of a concept, provided that:

- a concept can have several text entries;
- each of these text entries can have several senses in Wiktionary and in RuThes, these two sets of senses for a text entry can be different in size. For example, word *стрелка* is related to seven concepts in RuThes-lite and has eleven senses in Wiktionary.

  To extract a gloss for a given concept the following procedure was implemented:
- for all text entries of a concept, candidate glosses from Wiktionary are extracted. Glosses are cleaned from examples because examples can accidentally influence on matching. Then glosses are lemmatized, functional words are removed. So for every gloss we obtain the vector of lemmas.

For a concept we also create a vector. The vector includes all text entries of a concept, text entries of super-class concepts and whole-concepts. If a word is met several times in these text entries then its frequency in the vector is enhanced.

For example, concept *СТРЕЛКА РЕК* (*river spit*) has the following text entries and relations:

СТРЕЛКА РЕК (*river spit*)
(Syn: *стрека*, *стрелка рек*, *стрелка между реками*, *стрелочный*)
class: *КОСА* (*ВЫСТУП ЗЕМЛИ*) (Syn: *береговая коса*, *коса*, *коса берега*, *намывная коса*, *песчаная коса*)
asc1: *ВПАДЕНИЕ РЕКИ, ПОТОКА* (*stream inflow*) (Syn: *впадать*, *впадание*, *впадение*).

Therefore the following concept vector is generated for matching:
((*коса* 5) (*стрелка* 3) (*река* 2) (*стрелочный* 1) (*берег* 1) (*береговой* 1) (*намывной* 1) (*песчаный* 1) (*впадение* 1) (*впадать* 1) (*впадание* 1))

The relevant gloss from Wiktionary is as follows: "*узкий продолговатый участок суши, окружённый с трёх сторон водой, особенно на слиянии двух рек*". Its vector looks like:
((*узкий* 1) (*продолговатый* 1) (*участок* 1) (*суша* 1)
(*окруженный* 1) (*окружить* 1) (*особенно* 1) (*особенный* 1)
(*сторона* 1) (*вода* 1) (*слияние* 1) (*река* 1).

To this vector synonyms and hypernyms described in Wiktionary are added. In this case hypernyms *мыс* and *полуостров* are indicated for this sense in Wiktionary and therefore they are added with 1 count to the vector.

The matching weight between the concept vector and a gloss vector is equal to the scalar product of vectors without normalization, that is in the above-mentioned example the weight is equal 3 and this is the largest weight for all candidate glosses. As a result of this procedure around 60% of RuThes-lite concepts have obtained glosses.

Random testing of assigned glosses showed that 91% glosses were matched correctly to relevant concepts. Some glosses were missed, so recall is equal to 85%, F-measure—87.9%. In similar experiments on linking of GermaNet and Wiktionary,

authors of (Henrich et al., 2012) report that the best matching results (84.3%) were achieved using all relations of GermaNet with weights tuned for every type of relations. We did not use subclass relations (hyponyms) and parts, so some improvement of matching can be possible.

Currently, the list of extracted glosses is checked out by linguists, which remove irrelevant glosses and correct glosses with some problems of extraction.

4. Publication of RuThes

At present, RuThes thesaurus is partially involved in several commercial projects with other organizations and therefore it cannot be published as a whole. But the interest in a large thesaurus of Russian language is considerably growing therefore we decided to publish RuThes partially.

The first publicly available version of RuThes (RuThes-lite) is available from http://www.labinform.ru/ruthes/index.htm. We plan to distribute RuThes-lite as free for noncommercial use (Attribution-NonCommercial-ShareAlike 3.0 Unported license).

Conclusion

In this paper we presented RuThes linguistic ontology. This resource has been developed for a long time (more than fifteen years) and was used as a resource in various applications of NLP and information retrieval such as conceptual indexing, semantic search, query expansion, automatic text categorization and clustering, automatic summarization of a single document and multiple documents.

Now the first version of RuThes—RuThes-lite has been published. In this paper we described its structure and current state. We hope that this resource, having the broad and detailed lexical and terminological coverage of contemporary Russian news articles and official documents, will facilitate development of NLP techniques and research for Russian language.

In addition to publication of RuThes, we plan to automatically generate a resource in WordNet-like form (RuWordNet) including such relatively new information as WordNet domains and derivational links, which is widely discussed in the WordNet community. We think that RuThes contains enough data for generation such a resource. Its publication will be an important step in developing Russian semantic resources, connection with WordNet community.

Acknowledgements

The work is partially supported by Dmitrii Zimin Dynastia Foundation with financial support of Yandex founders.

References

1. *Agirre E., Soroa A.* (2009), Personalizing pagerank for word sense disambiguation, Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.
2. *Azarowa I.* (2008), RussNet as a Computer Lexicon for Russian, Proceedings of the Intelligent Information systems IIS-2008, pp. 341–350.
3. *Azarova I., Mitrofanova O., Sinopalnikova A., Yavorskaya M., Oparin, I.* (2002), Russnet: Building a lexical database for the Russian language, Proceedings of Workshop on Wordnet Structures and Standardisation and How this affect Wordnet Applications and Evaluation, Las Palmas, pp. 60–64.
4. *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.
5. *Baccianella S., Esuli A., Sebastiani F.* (2010), SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining, Proceedings LREC-2010, Vol. 10, pp. 2200–2204.
6. *Bentivogli L., Forner P., Magnini B., Pianta E.* (2004), Revising WordNet domains hierarchy: semantics, coverage, and balancing, Proceedings of COLING 2004, Geneva, Switzerland, pp. 101–108.
7. *Bhatt B., Kunnath S., Bhattacharyya P.* (2014), Graph Based Algorithm for Automatic Domain Segmentation of WordNet, Proceedings of Global WordNet Conference GWC-2014.
8. *Braslavski P. I., Mukhin M. Y., Lyashevskaya O. N., Bonch-Osmolovskaya A. A., Krizhanovsky A. A., Egorov P.* (2013), Yarn Begins. http://www.dialog-21.ru/digests/dialog2013/materials/pdf/BraslavskiyP_YARN.pdf
9. *Clark P., Fellbaum Ch., Hobbs J.* (2008), Using and Extending WordNet to Support Question-Answering, Proceedings of Fourth Global WordNet Conference (GWC'08), Hungary, Szeged, pp. 111–119.
10. *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.
11. *Guarino N., Welty Ch.* (2000), Ontological Analysis of Taxonomic Relationships, In Conceptual Modeling (ER-2000), Springer, Berlin Heidelberg, pp. 210–224.
12. *Fellbaum Ch.* (1998), WordNet: An Electronic Lexical Database, Cambridge, MA, MIT Press.
13. *Henrich V., Hinrichs Er., Vodolazova T.* (2012), Semi-Automatic Extension of GermaNet with Sense Definitions from Wiktionary, Proceedings of LREC-2012.
14. *Loukachevitch N., Dobrov B.* (2004), Sociopolitical Domain as a Bridge from General Words to Terms of Specific Domains, Proceedings of Second International WordNet Conference GWC-2004, Brno, pp.163–168.
15. *Loukachevitch N.* (2009), Concept Formation in Linguistic Ontologies, Conceptual Structures: Leveraging Semantic Technologies. Proceedings of ICCS-2009, Springer Verlag, LNAI-5662, pp. 2–22.

16. *Loukachevitch N.* (2011), Thesauri in information-retrieval tasks, Moscow, Moscow University publishing house.
17. *Loukachevitch N., Dobrov B.* (2014), RuThes Linguistic Ontology vs. Russian Wordnets. Proceedings of Global WordNet Conference GWC-2014, Tartu.
18. *Koeva S., Krstev C., Vitas D.* (2008), Morpho-semantic relations in Wordnet–a case study for two Slavic languages. In *Proceedings of the Fourth Global WordNet Conference*, pp. 239–254.
19. *Pedersen T., Patwardhan S., Michelizzi J.* (2004), WordNet: Similarity: measuring the relatedness of concepts, Demonstration Papers at HLT-NAACL 2004. Association for Computational Linguistics.
20. *Z39.19.* (2005), Guidelines for the Construction, Format and Management of Monolingual Thesauri. NISO.

DESIGNING "HUMAN CHARACTERS" LEXICAL DATABASE

Lukashevich N. Ju. (natalukashevich@mail.ru),
Kobozeva I. M. (kobozeva@list.ru)

Lomonosov Moscow State University, Moscow, Russia

The paper discusses a general layout of "Human Characters" lexical database specifically developed to study the meanings of words from the semantic field of human character traits. It is intended as a resource providing a format for a comprehensive analysis of character words usage in different languages. A database with contexts from large modern corpora is considered a convenient tool for semantic analysis which offers such advantages as facilitating data storage and presentation, and keeping the analysis consistent while making changes possible at the same time. It is shown how several issues which significantly influence the analysis procedures are resolved in the pilot database version. These include identifying relevant contexts, describing features of a typical situation in which the character trait in question is exhibited, and comparing contextual meanings of the studied words. The suggested technique provides a more flexible tool for capturing similarities and differences between contexts within one language on the one hand, and gives ground for comparing the usage of translation equivalents on the other.

Key words: cognitive semantics, database, semantic analysis, human character, corpora contexts, behaviour pattern

Our work is devoted to the analysis of words from the semantic field of human characters—words either naming a person in accordance with the person's character traits (e.g., *greedy*, *avaricious*, *mean*; *meanie*, *cheapskate*) or the traits themselves (*greed*, *avarice*), or their derivatives (e.g. *meanly*, *greedily*; Russian *zhadnichat'* 'to be greedy' derived from *zhadny* 'greedy').

The semantic field of human characters

In psychology where the phenomenon of human character is the subject of study, there exists a variety of theories of human character and personality. In Russian psychology human character is mostly regarded as a system of individual psychological features revealed in human behaviour (in various types of activity, communication and interaction with other people). A character trait is seen as a stereotype of behaviour which is realised with high probability in a relevant situation. It is generally accepted that volitional qualities constitute a part of human character, whereas intellectual qualities do not. The relationship between character and temperament is still a subject of discussion.

To identify the limits of the semantic field of character one should not rely on scientific notions because language reflects naïve psychology: we need to know what properties speakers of a language consider as pertaining to human character (if this language has a lexical item with a given meaning). To do it, a special experiment was designed in (Lukashevich 2004) to check whether clear-cut temperamental, intellectual and volitional qualities represent character for native Russian speakers. According to the results obtained, volitional qualities are, whereas intellectual qualities are not regarded as character traits by Russian speakers. As for temperament and character, there is no clear boundary between these notions in Russian (for more details see (Lukashevich 2004)).

The nature of semantics of words from this field has been previously discussed in (Lukashevich 2002, Lukashevich 2004). Character words pose problems for a lexicographer, which is why their meanings are poorly represented in existing dictionaries (Kobozeva, Lukashevich 2012). A more effective approach was introduced in (Lukashevich 2004) It takes into account the above mentioned link between a character trait and the situation in which it is usually triggered. According to this approach a character word meaning is represented with the help of behaviour pattern—a generalised implicative scheme which links the initial typical situation with the stereotyped behavioural response of a person with this trait of character. (The notion («shablon povedenija» in Russian) was first proposed in (Martemianov, Dorofeyev 1969) and discussed in detail in (Martemianov 1999, Lukashevich 2002, Lukashevich 2004)).

To obtain the information about the typical triggering situation and the actions performed by a person characterised by a character word a thorough and extensive study of this word's use is required. This may be done by analysing large numbers of contexts with the word (thus possibly repeating the work a native speaker does when acquiring such notions). A database containing contexts from large modern corpora can help to keep the analysis consistent, and provide a convenient way to store both the material for studying and the results of such analysis.

Words denoting human character in existing database projects

Taking a look at the existing DB projects one can say that they are mostly not well suited for the purpose of performing an in-depth analysis of character word meanings. An important feature of this field is that it is mostly represented by rows of near-synonyms. The way they are handled in the current versions of such projects as WordNet and FrameNet has already been discussed in (Lukashevich, Kobozeva 2011): though both resources provide valuable information (WordNet—on synonymic relations between lexemes in a language[1], FrameNet—on the roles played by participants of the described situation), they leave the differences between near-synonyms mostly unclear. FrameBank, a Russian-language project of the FrameNet type, is intended as a hybrid of a dictionary of constructions with an annotated corpus. Another

[1] WordNet related projects, such as Open Multilingual WordNet, can even relate synsets in various languages to each other

Russian-language project called Lexicograph mainly studies the relations between different meanings of a word and how each meaning predicts the way the word is used (and covers only verbs in its current version). As for the Typological database "The Vocabulary of Pain" it was developed to study a specific thematic class of words across languages, and its design, though remarkably flexible (Kostyrkin et al. 2012), does not suit the features of words from our semantic field (as what is studied in our case is not always explicitly present in the text).

In general it should be noted that in most of these projects the semantic field of human character is rather poorly covered (if intended to be included at all) (for example, such synonymic row as "greedy, covetous, avaricious" is not present in the current version of FrameNet). More importantly, the above mentioned projects seem more focused on the meaning of a word as a whole undergoing various processes like meaning shifts, and none intend to go deeper into the meaning itself. This research aims (beside a major goal of building a more adequate semantic representation of character words by identifying their behaviour patterns as well as their prototypical ("best") examples) to reach such a level of detail where distinctions between near-synonyms will be visible.

All of this called for developing a DB specifically designed for the purposes of the present project.

"Human Characters" Database: major problems

The present paper discusses a general outlay of "Human Characters" database (in its pilot version)[2]. It is intended as a resource providing a format for comprehensive analysis of character words usage in different languages[3]. The DB is supposed

[2] The work is being done within the framework of a seminar "Human characters through the prism of language" at the Department of Theoretical and Applied Linguistics, Philological Faculty, Moscow State University. At the moment it encompasses examples from the main subcorpus of the Russian National Corpus on several Russian adjectives and nouns from Greediness and Candidness semantic groups. The initial analysis of contexts is mostly done by students who fill in the DB in accordance with the provided guidelines After that everything is reviewed by the developers (i.e. the authors of the present paper), which means that every context is analysed at least by three persons (all native Russian speakers). Those issues which cause disagreement are then discussed at the seminar. If anything still remains unclear at this stage, such issues may be tested through specially designed experiments with native speakers (as in (Lukashevich 2004)).

[3] Some time ago contexts from the British National Corpus displaying the use of several English adjectives and adverbs from Candidness group were taken for preliminary analysis. Some results of this analysis were presented in (Kobozeva, Lukashevich 2012). However, it became clear that BNC does not contain enough contexts to make reliable conclusions about fine distinctions between near-synonyms in a row. Besides, extracting all the relevant contexts from BNC presents certain technical problems, which means that it will be necessary to use material from other (bigger) corpora of English. Another crucial point will be to ensure that the results are reviewed by native speakers. (It should be noted here that the examples in the paper are given not only in Russian, but also in English on purpose. This is done to avoid unnecessary translation (in cases where a similar English example was easy to give) and to show that at least some features are shared by words from this group across languages.)

to list all the contexts in which the studied word is used in a given language corpus. Each (relevant) context is described with great detail from various angles: not only the exact meaning of the word in question, but most varied aspects of its usage and features of the context are identified (like the grammatical construction in which it is used, whether it conveys any evaluative meaning, whether there is a reference to a fiction character, etc). By generalising from contextual details in every aspect one can identify the major features of a typical situation associated with the named character trait and then formulate its behaviour pattern (i.e. specify which actions in what conditions are usually performed by a person so that this person may be assigned the named character trait).

Thus, for a comprehensive study of a character word various aspects of its usage should be taken into account. It should be remarked that not all these aspects and consequently not all details of the DB structure will be discussed here, but only those which significantly influence the analysis procedures.

1. Identifying relevant contexts

First of all, an important task is to identify the contexts which are relevant for this analysis. The thing is that it is typical for many adjectives and adverbs belonging to this group to have more than one meaning, or at least more than one usage. Character adjectives may be used with nouns (both animate and inanimate), as well as in predicative constructions. For example, *candid* can be found in the following contexts:

(1) a) *Maybe I delivered my opinion more bluntly than I should have, but I had always been candid with Ted.*
b) *He is either a* ***candid*** *friend or an honest enemy who disdains to tell lies.*
c) *…her femininity was exquisitely* ***candid****.*
d) *…his* ***candid*** *eyes never left Pitt's.*
e) *We take a* ***candid*** *look at the choices now open to you.*
f) *We are doing a* ***candid*** *camera in here today.* (the British National Corpus (BNC))

Of the above only (1a) and (1b) describe a person and their actions directly and can be used to find explicit information related to why the person was assigned the characteristics in question. Although other examples may also be useful on further stages of study, especially when fine distinctions between near–synonyms are elucidated, such contexts are not included into the "core" analysis.

Character nouns can also display certain ambiguity, though this is not typical of the "core" members of the semantic field (derived from character adjectives—*zhadina* 'cheapskate' from *zhadny* 'greedy' in Russian, *meanie* from *mean* in English). But there are less central cases of metonymic or metaphoric nominations derived from names belonging to other semantic classes. For instance, *lisa* 'fox' in Russian has two meanings: 1) 'an animal'; 2) 'a cunning, honey-mouthed person' (Ozhegov 1990). For the purposes of this research only such contexts are considered relevant where the word in question either characterises a person in accordance with the

person's behaviour, or metonymically signifies some aspects of behaviour typical for a person with a particular character trait.

As for character adverbs, they are also used in more than one way, and not all of their uses need to be taken into account for the purposes of this research (Kobozeva, Lukashevich 2012).

This means that it is important for a researcher to be able to sort corpora contexts in accordance with their relevance for analysis. (It should be noted here that at the moment all the contexts which are extracted from a corpus in accordance with the search conditions are included into the DB irrespective of their relevance for several reasons. On the one hand, it is highly problematic to sort contexts automatically, so that only relevant ones are imported into the DB. On the other hand, the contexts which are not relevant for the "core" analysis may prove necessary on later stages of study, as it has already been mentioned above. Besides, the information they provide may be used for other purposes (e.g., to analyse regular polysemy models).)

Defining grammatical construction is helpful, but it is not enough to separate all the relevant contexts. For example, adjectives in predicative constructions can relate both to animate and inanimate nouns, and only sentences with animate nouns would be normally relevant here. On the other hand, (2) shows an example where personification takes place and (2) should therefore be taken into account:

(2) *The bathroom mirror was **candid**, almost disapproving, whereas her bedroom mirror took and returned a more indulgent view, softening lines and contours.* (BNC)

To resolve this, a special field "Relevant example" was introduced, so that when marked it signals that the context is relevant for analysis. (Nevertheless grammatical construction still needs to be identified, as these distinctions may prove significant for the choice of contextual meaning.)

2. Context length

It should be noted that whenever it is possible, it is important to have at hand not only sentences with the words in question, but longer contexts (as long as can be extracted from the corpus). The reason is that quite often the information necessary to specify the meaning in which the word is used is contained elsewhere in the text (mostly within the range of several sentences before or after the sentence with the word (e.g. as in 3a below)). (Here it should be noted that this is relevant only for those corpora which allow extracting more than one sentence with the word used, like the Russian National Corpus (RNC), for example. Some resources, such as the BNC, do not allow this, which often makes examples useless for the semantic analysis we perform.)

As quite lengthy pieces of text are required sometimes to understand the situation represented in the context (i.e. to be able to identify the conditions and the action of the person characterised by the word in question), a special field "Pattern Instance" (i.e. a specific realisation of a particular behaviour pattern linked with a particular

character trait) was introduced in the DB to make analysis and further discussion easier: short descriptions, or synopses, of the situations from the contexts should be recorded there. The general idea is to get an additional intermediate level of generalisation which would allow working with the example without having to read the whole text once again every time. These short descriptions will contain example information in a form which will make it easier to see in what way the given situation instantiates behaviour pattern. Such paraphrases should keep all the relevant features of the example and yet be short and concise. They should include only essential details of the situation, such that if these are left out, the example becomes unclear, or it is no longer enough to assign the named character trait to a person. (3a) and (3b) below show an English translation of a sample context from RNC illustrating the use of *korystolyubivy* 'avaricious' and a resulting Pattern Instance :

(3) (a) … *the Russian revolution brought … the petty bourgeoisie to the front row. Yes indeed, the very crass,* ***avaricious*** *petty bourgeoisie, which back at the beginning of our century was ridiculed by all liberal Russian writers, from Chekhov to Gorky. … It was exactly the petty bourgeois—selfish, apolitical and devoid of ideology—who had waited out in the backwater for the Civil war storms to calm down, to then crawl out safe and sound and serve the Soviet power; it did not matter what kind of power to serve—just power, in order to grab a piece of the government pie as big as possible*[4]. (the Russian National Corpus—Anatoly Gladilin, A long race day (1976–1981)[5])
(b) Pattern Instance: *lower middle class waited out the revolution and the civil war and began to serve the current power* (no matter which one, here—the Soviet power) *to get as much money, property, etc. as possible*

3. Defining features of a typical situation

Another important task is to find the defining features of a typical situation in which the character trait in question is revealed. Initially the plan was to identify the Action, Motivation and briefly describe the overall Conditions of the situation (for example, just to indicate that it concerned "paying somebody for the work done"). However, it was decided at a certain point that it would be preferable to identify roles or participants of the situation in every relevant context as this information may give

---

[4] … русская революция выдвинула на авансцену… мещанство. Да, да, то самое дремучее, **корыстолюбивое** мещанство, которое еще в начале нашего века осмеивали все прогрессивные русские писатели, начиная с Чехова и кончая Горьким. … Именно мещанин, эгоистичный, аполитичный и безыдейный, переждал в тихой заводи, когда успокоятся бури Гражданской войны, и целым и невредимым вылез наружу, чтобы служить советской власти, не важно какой, важно, что власти, чтобы отхватить себе кусок правительственного пирога побольше. (НКРЯ — Анатолий Гладилин. Большой беговой день (1976–1981))

[5] Authors' translation

clues to the differences between near-synonyms[6]. Thus, such fields as Subject, Object-Theme and 2nd Participant were introduced: Subject defines a participant of the situation which is characterised by the word in question; Object-Theme names such entity the attitude to which serves as the basis for distinguishing the character trait (or, to be more precise, the cluster of character traits); and 2nd Participant describes the participant whose interests are affected by the Subject's behaviour. For example, in (3) "lower middle class" is the Subject, "material wealth" is the Object, and 2nd Participant is missing. The Action in this example is "serve the current power", Motivation is "to obtain as much material wealth as possible", Conditions may be defined as "when there is a choice between obtaining material wealth and following one's political principles".

4. Representing the meaning

The third task constituting the crucial part of analysis is to find differences and similarities between the exact meanings in which the words in question are used in contexts.

Initially a sort of a two-level tree structure was used to make such generalisations. The main point was to get non-overlapping subsets of contexts and to specify what makes the cases of human behaviour similar within each subset. This helped to identify the main features of actions typical for various instances of behaviour associated with the studied character trait. As it was merely an intermediate step, it was not of much consequence that the resulting groups of contexts were not always of "equal weight", e.g. not always on the same level of generalisation. For example, for Russian *otkrovenny* ('frank,candid') the groups were as follows:

1) the person says smth about him/herself which is not obvious and this may lead to negative consequences to the person:
   - tells smth negative about him/herself;
   - speaks of his/her real feelings;
   - says what (s)he is thinking;
2) the person does not hide anything;
3) the person says what (s)he is thinking although it violates etiquette rules;
4) the person tells the listener smth negative about the others which is not obvious; etc.

Thus it was clear that for *otkrovenny* a second level of generalisation could be identified as several meanings shared a common part.

However, one level of generalisation could also be enough, as was the case for Russian *iskrenny* ('sincere') where the groups were as follows:

1) the person says what (s)he is really thinking;
2) what the person is doing, corresponds to what (s)he is saying and thinking;

---

[6] Here not the semantic-syntactic roles of the sentence with the character word are meant, but roles of the participants of the behaviour pattern.

3) the person says what (s)he is really thinking and this leads to negative consequences for others;
4) the person says what (s)he is really thinking and this leads to negative consequences for this person;
5) the person demonstrates feelings, which (s)he is really experiencing;
6) the person says smth about his/her real feelings and thoughts which is unpleasant for the listener; etc.

It is easy to note that certain parts of the meanings roughly formulated above are repeated in various combinations. The way they are combined does not allow to fit them into a clear tree structure. However, the resulting picture is of much help in identifying differences between these two near-synonyms and making higher level generalisations for behaviour patterns.

Taking all of this into account a two-level structure was introduced in the DB consisting of Meaning General and Meaning Subtype. The fields were supposed to divide the contexts into non-overlapping subsets, such that each context within a subset depicts one type of situation sharing some features and distinctly different from situations in other subsets. Meaning Subtype is supposed to specify Meaning General, in a way that Meaning General is subdivided into several Meaning Subtypes.

However, an attempt to analyse in a consistent way large numbers of examples with several words belonging to the same group (Greediness) showed that this would not work as intended. Firstly, it proved quite difficult to get non-overlapping groups (no matter how small they were). Secondly, there were cases when some elements of sense appeared sporadically and their presence did not seem to be linked to any particular Meaning General or Meaning Subtype. (For example, such was the pleasure a person felt when in direct physical contact with money he possessed—a sense identified in several contexts for *skupoy* 'stingy' and *zhadny* 'greedy' in Russian.)

Thus it soon became clear that though still being helpful, this two-level structure is not enough for capturing the distinctions identified.

The above mentioned difficulties with this group of words were only to be expected due to such feature of their semantics as Wittgensteinian family resemblance (a situation when objects in a set share common features in pairs, but no object has all the features at once (Wittgenstein 1953)). Family resemblance effects have already been mentioned as the reason why it is difficult to identify the boundaries between rows of synonyms and between volumes of near-synonyms meanings in a row (Lukashevich 2004, Kobozeva, Lukashevich 2011). The above mentioned difficulties show that similar effects are seen not only on the level of boundaries between the meanings of words, but also between various usages of one word.

Understandable and foreseeable as these difficulties are, it was necessary for the research purposes to find some way around them as this interim step is a key part of the analysis. Obviously this is yet another example of non-discreteness in language which has to be accounted for using discrete tools (in the absence of any other ones which would better suit the nature of the material), a problem discussed in (Kibrik 2013).

To overcome them we use the following technique: we introduce a list of component elements of meaning and in each context choose the ones relevant to this

particular instance. These elements are such "chunks" of meaning which have been identified on the previous stages of analysis as parts which are used as whole units. For example, it is clear in the roughly named meaning groups for *otkrovennny* and *iskrenny* that a part of meaning 'what X is saying may lead to negative consequences to X' may be added to other "blocks" like that and this way form a meaning in which the character word is used. When analysing examples with words which belong to the Candidness group in English (*candid, frank, sincere, open*) it was obvious that this "block" is also often present in their meanings, as well as other "blocks" of meanings which can be identified for *otkrovennny* and *iskrenny* (Kobozeva, Lukashevich 2011). Therefore it seems possible to make a list of such parts of meaning for words instantiating the Candidness frame (in the FrameNet sense) and use it to describe various usages of words belonging to this group (possibly not only in English and Russian). Thus, obtained through a procedure similar to the one used by A. Wierzbicka when identifying semantic primitives (Wierzbicka 1972), such components of meaning represent the level on which it is possible to compare the meanings of near-synonyms in a row in one language as well as the meanings of translation equivalents across different languages.

In the course of analysing Russian adjectives and nouns from the Greediness group the following list was identified:

- seek not to spend resources
- seek not to spend money on others
- seek to increase their material wealth
- seek to obtain material wealth
- attach primary importance to money
- attach primary importance to material values
- violate moral (=ethical) rules with regard to others
- violate legal norms
- seek to be the sole user of a resource
- commit an action undesirable for its object/experiencer
- etc.

At the moment the list includes 27 elements. There will most likely be additions when greediness in other languages comes under analysis, however, the "core" list presumably will not change greatly. It is worth mentioning that such components of meaning should be expressed using a semantic metalanguage (in a sense that they should contain units from a final list, such units should be used in a consistent manner, and their meanings may not be identical to the meanings of these words when used in a real language)[7].

From the example above it is clear that such elements will most likely have different "weight" in the list: some will be more central for a particular group (like the ones related to the attitude to material values for the Greediness group), whereas others, more "peripheral" ones, will include concepts from other spheres (e.g. 'violate moral rules with regard to others' is relevant not only for *korystolyubivy* 'avaricious' but also for *verolomny* 'deceitful', etc.).

[7] In this project metalanguage is being built incrementally, one character group after another.

Whether such lists can be universal requires further research, but they will at least provide a more flexible tool for capturing similarities and differences between contexts within one language, and at the same time give a ground for comparing the usage of translation equivalents on the other.

To illustrate what has been said above we show how semantic information for (3a) and (3b) may be recorded in the DB:

(level 2) Meaning Subtype = *the Subject attaches great importance to material values and seeks to obtain material wealth using his/her proximity to power*

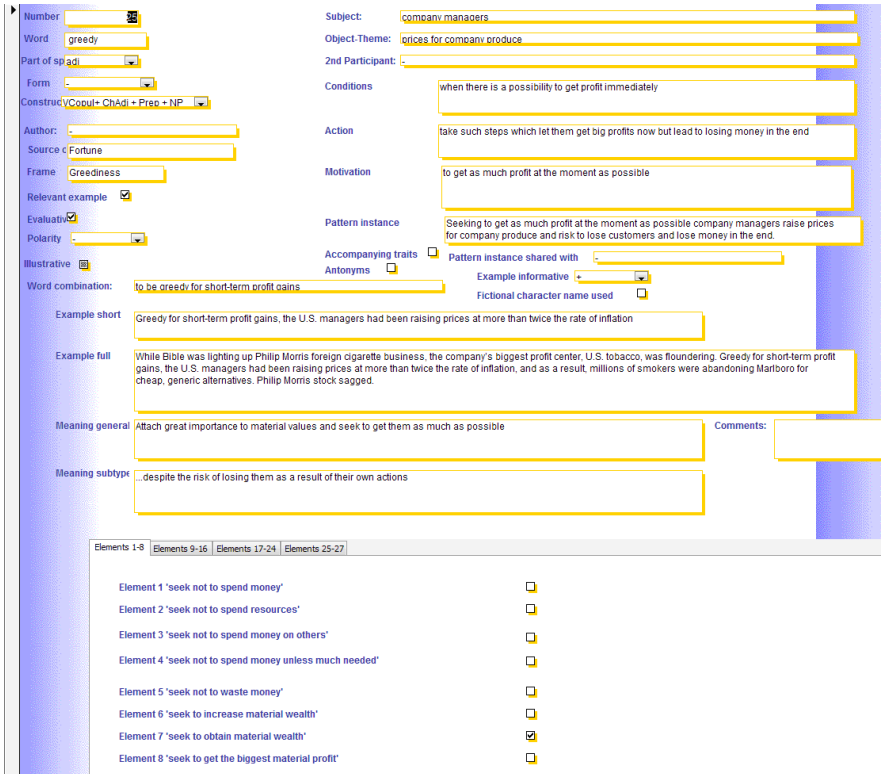
Components of meaning: *'attach great importance to material values' + 'seek to obtain material wealth' + 'use their proximity to power' + 'as much as possible'*

(level 1) Meaning General = *attach great importance to material values and seek to obtain material wealth violating, in so doing, certain rules*

(level 0) Behaviour Pattern (tentative):
Korystolyubivy =
= *when obtaining material wealth is possible such people prefer material values over other values and seek to obtain or increase their material wealth though it may lead to negative consequences for them and/or other people.*
It is typical for such a person:

- *to choose a job offer with the highest payment (when it is not necessary) though it may force them and their family to live in extreme conditions;*
- *to serve no matter which power so that they can use their position to get money, property, etc.;*
- *to betray a friend for money;*
- *for a doctor to refuse to go and see the patient for the second time if (s)he does not get paid for the visit;*
- *for an official to take bribes.*

The DB also contains fields describing other important aspects of use (such as the presence of an evaluative component) or important features of the context (e.g., the presence of synonyms in conjunction with the word, etc.), which will not be discussed here. To give an idea of what it looks like, Fig. 1 shows a print-screen from the DB with an example from the Corpus of Contemporary American English. It is important that the suggested DB should allow to use the various recorded analysis results for the purposes other than the main research goals (e.g. use the information on evaluative component for sentiment analysis).



References

1. British National Corpus, available at http://www.natcorp.ox.ac.uk/ (as of May 2012)
2. Corpus of Contemporary American English, available at http://corpus.byu.edu/coca/ (as of 20.04.2014)
3. FrameBank, available at http://framebank.ru/ (as of 20.04.2014)
4. FrameNet Project (II), available at https://framenet.icsi.berkeley.edu/fndrupal/
5. *Kibrik A. A.* (2013) Non-discreteness in language and focal structure [Nediskretnost' v jazyke i fokalnaja struktura], Cognitive Modelling: Proceedings of the 1st International Forum on Cognitive Modelling [Kognitivnoe modelirovanie: Trudy Pervogo Mezhdunarodnogo foruma po kognitivnomu modelirovaniju], Italy, Milano-Marittima, part 1, pp. 113–116.
6. *Kobozeva I. M., Lukashevich N. Ju.* (2012), Human characters through the prism of adverbs, 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, pp. 277–287.

7. *Kostyrkin A. V., Panina A. S., Reznikova T. I., Bonch-Osmolovskaya A. A.* (2012), Constructing a lexico-typological database (for a study of pain predicates), 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, pp. 288–295.
8. Lexicograph Project (2010), available at http://lexicograph.ruslang.ru/ (in Russian)
9. *Lukashevich N. Ju.* (2002), Predicates describing character traits and behaviour patterns [Kharakterologicheskie predikaty i shablony povedeniia], Journal of MSU, Philology series [Vestnik MGU, ser. Filologiia], Moscow, 2002. № 5, pp. 131–141.
10. *Lukashevich N. Ju.* (2004), Cognitive semantic analysis of predicates denoting human character traits [Kognitivno-semanticheskii analiz predikatov, oboznachaiushchikh cherty kharaktera cheloveka], PhD thesis, Moscow, 2004.
11. *Lukashevich N. Ju., Kobozeva I. M.* (2011), Character nominations in ontological perspective, Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2011" [Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Po materialam ezhegodnoi Mezhdunarodnoii Konferentsii "Dialog 2011"], Bekasovo, pp. 468–477.
12. *Martem'ianov Ju. S.* (1999), Metalanguages of sentence and text description [Metaiazyki opisaniia predlozheniia I teksta], Text processing and cognitive technologies [Obrabotka tekst i kognitivnye tekhnologii], Moscow, 1999. № 3, pp. 124–129.
13. *Martem'ianov Ju. S., and Dorofeev G. V.* (1969), Logical inferences and eliciting relations between sentences in the text [Logicheskii vyvod i vyiavlenie sviazei mezhdu predlozheniiami v tekste], Machine translation and applied linguistics [Mashinnyi perevod i prikladnaia lingvistika], Moscow, 1969, issue 12, pp. 36–60.
14. Open Multilingual WordNet (1.0), available at http://compling.hss.ntu.edu.sg/omw/cgi-bin/wn-grid.cgi
15. *Ozhegov S. I.* (1990), Dictionary of the Russian language [Slovar' russkogo jazyka], Russkij jazyk, Moscow.
16. Russian National Corpus, available at http://www.ruscorpora.ru/ (as of 01.2014)
17. Typological database "The Vocabulary of Pain" available at http://orientling.ru/bolit/ (as of 20.04.2014),
18. *Wierzbicka A.* (1972), Semantic Primitives, Athenäum-Verlag.
19. *Wittgenstein L.* (1953), Philosophical Investigations, Blackwell Publishing. 2001.
20. WordNet Lexical Database (version 3.1), available at http://wordnetweb.princeton.edu/perl/webwn (as of 20.04.2014)

ОЦЕНКА РЕЗУЛЬТАТОВ ПАРСЕРА: РАСПОЗНАВАНИЕ СЕМАНТИЧЕСКИХ РОЛЕЙ УЧАСТНИКОВ ФРЕЙМОВ В ЯЗЫКЕ С ПАДЕЖНЫМ МАРКИРОВАНИЕМ

Ляшевская О. Н. (olesar@gmail.com)

Национальный исследовательский университет
Высшая школа экономики, Москва, Россия;
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

В статье обсуждаются подходы к оценке парсеров, задачей которых является автоматическое определение семантических ролей (semantic role labeling, SRL). Как было показано ранее, качество распознавания именованных семантических ролей в стиле FrameNet в большой степени зависит от количества выделяемых ролей и может падать, если инвентарь ролей в ресурсе, используемом для обучения, и инвентарь ролей в целевом ресурсе разнится. Наше исследование представляет первый шаг к созданию системы 'умной' оценки SRL-парсеров, которая вводила бы лингвистически мотивированные критерии оценки работы SRL-системы; позволяла бы классифицировать ошибки от незначительных до критически важных; была бы устойчива к возможным расхождениям между инвентарями ролей.

Статья описывает эксперимент, материалом для которого служит база данных FrameBank—общедоступный онлайн-ресурс, идеологически связанный с системой FrameNet и объединяющий словарь лексических конструкций частотных русских глаголов и размеченный корпус их реализаций в примерах из НКРЯ. Одним из параметров разметки аргументов конструкций служат их семантические роли, инвентарь которых в системе FrameBank устроен иерархически и представлен в форме графа. Исследуются статистические критерии дистрибуции ролей в словаре конструкций и расположение ролей на графе для того, чтобы сопоставить ответ системы и ответ золотого стандарта.

Ключевые слова: конструкции, семантические роли, полисемия, автоматическое определение семантических ролей, корпусная лингвистика, лексические ресурсы, эвалюация парсеров

EVALUATION OF FRAME-SEMANTIC ROLE LABELING IN A CASE-MARKING LANGUAGE

Lyashevskaya O. N. (olesar@gmail.com)
National Research University
Higher School of Economics, Moscow, Russia;
Vinogradov Russian Language Institute
of the Russian Academy of Sciences, Moscow, Russia

Kashkin E. V. (egorkashkin@rambler.ru)
Vinogradov Russian Language Institute
of the Russian Academy of Sciences, Moscow, Russia

The paper discusses evaluation techniques for semantic role labeling in Russian. It has been shown that the quality of FrameNet-style semantic role labeling largely depends on the quantity of roles and may decrease if the inventory of roles in the training set differs from that in the output resource. Our study is the first step towards the 'smart' evaluation tool which would introduce linguistically relevant criteria to evaluation; be able to put the mistakes on a scale from minor to critical ones; make evaluation easier in case the grid of roles varies.

We run an experiment based on the data from the Russian FrameBank, a FrameNet-oriented open access database which includes a dictionary of Russian lexical constructions and a corpus of tagged examples. The semantic role is one of the parameters that define the predicate-argument patterns in FrameBank. The inventory of roles is modeled hierarchically and forms a graph. We explore the cases when the role induced by the system and the answer of the gold standard do not match. We analyze the statistical criteria of distribution of roles in the patterns and the distance between the source and the target in the graph of roles as a mean to assess the goodness of fit.

Keywords: constructions, semantic roles, polysemy, semantic role labeling, corpus linguistics, lexical resources, evaluation

1. Background

Syntactic parsing and semantic role labeling (SRL) are two closely related tasks employed in the shallow semantic understanding of natural text. SRL focuses on the automatic identification and labeling of the relations between a predicate and its arguments which involves generalization over surface (morpho)syntactic patterns. It is a further step towards finding projections of semantic arguments in syntactic structures. With the advent of large-scale annotated data resources such as treebanks, PropBank (Palmer et al. 2005) and FrameNet (Fillmore et al. 2003), both domains have recently benefited from an enormous boost in machine learning methods. What

matters even more is the development of standard test data sets and evaluation metrics such as CoNLL 2007, CoNLL 2008 and SemEval 2007.

SRL can be divided into the following steps:

- Step 0. Target predicates (or frame-evoking words) are marked in the sentence.
- Step 1. Each target is disambiguated to a particular sense or semantic frame.
- Step 2. Words in context are classified into arguments and non-arguments; if a dependency tree is available, nodes are classified into actants and circonstants (in Tesnière 1959's tradition), i. e. 'inner arguments' and 'free modifiers'.
- Step 3a. The arguments are labeled as ARG0, ARG1, ARG2, etc. (PropBank-style SRL).
- Step 3a. The arguments are labeled with particular frame-relevant roles such as Agent, Experiencer, Stimulus, Path, etc (FrameNet-style or 'deep' SRL).

SRL is usually constrained to the target's locally expressed semantic arguments, i. e. syntactic dependencies. More advanced tasks, such as finding and resolving null instantiations from the surrounding context (Gorinsky, Ruppenhofer 2013); finding new edges introduced by the semantic structure, are currently out of the scope of industrial standards (see Das et al. forthcoming, Màrquez et al. 2008, Palmer et al. 2013 for a comprehensive overview).

Most SRL parsers stop after Step 3a. They solve the classification task for up to 10 clusters and there is an excess amount of training data on hand to reach good results. CoNLL 2008 shared task training and test data set (Surdeanu et al. 2008) provides a standard benchmark.

FrameNet-style SRL presupposes identifying targets that could evoke frames in a sentence, identifying the correct semantic frame for a target, and finally determining the arguments that fill the semantic roles of a frame. The parsers of this type use much more fine-grained structure of clusters as an input. SemEval 2007 benchmark data set (Baker et al. 2007) provides 665 labels whereas FrameNet 1.5 release has as much as 877 labels, so the optimization of verb classes and semantic roles clusters is considered helpful to overcome the sparse data problem. Other related tasks include discovering new semantic frames and roles, i.e. associating frames to 'unseen' lexical items which cannot be found either in FrameNet or in training data.

Both PropBank-style and FrameNet-style SRL tasks are language dependent. We can assume that such factors as left/right position against the predicate, voice, lexical and semantic cues, case and preposition marking, the general surface obligatoriness of arguments, would have uneven impact, depending on the language. Yet a more important factor seems to be the amount of corresponding annotation in training resources available across languages. Hajič et al. 2009 show that if a SRL tool is applied to different languages, its performance can drop by 10% (e.g. from F1≈85.5 for English to F1≈76.5 for Japanese and Spanish).

The objectives of this paper are to propose a benchmark and evaluation scenario for Russian frame-semantic role labeling. We target two well-known problems in parser evaluation: the comparability of output role labels and insufficiency of traditional performance measures (precision P, recall R, F1) in evaluation against the gold standard. The paper is organized as follows. Section 2 outlines the design and evaluation metrics. We argue that if the standard list of roles is connected into a graph, this

can help assess the SRL results which otherwise may be difficult to compare. In Section 3 we introduce FrameBank as an open resource that can be used in SRL training and/or evaluation based on Russian data. Section 4 summarizes an experiment on SRL for Russian prepositional phrases, including the structure of the data used, the rules and the qualitative analysis of the results. In Section 5 we come back to the metrics proposed earlier and show the evaluation scenario at work.

2. SRL Evaluation Metrics

The standard approach to NLP evaluation assumes that there exists a test corpus provided with a 'Gold Standard' (GS) annotation. Let G = {s^g_1, s^g_2, … s^g_N} be a set of semantic roles in the GS. Given the output from an NLP tool E = {s^e_1, s^e_2, … s^e_N}, we can compare it against the GS set and compute the number of matches M with respect to the number of answers E returned by the parser (i.e. precision P = #M / #E), the number of matches M with respect to the total number of elements G in the GS (i.e. recall R = #M / #G), and their harmonic mean F-score.

But what if a parser is either developed in a different framework or trained on a different data set, or trained on unlabeled data? RU-EVAL evaluation forum[1] has shown that many Russian parser developers rely heavily on the size and quality of their own training resource. If we project this to the domain of SRL, we can expect that the inventories of possible answers (semantic roles) in the SRL resources might vary significantly (cf. Azarowa 2008; Ermakov, Pleshko 2009; Petrova 2013; Smirnov, Shelmanov 2014; Kashkin, Lyashevskaya 2013, among others), what would make the comparison not straightforward.

Lang and Lapata (2011) suggest another set of evaluation metrics that assess an overall goodness of clustering and can work with unsupervised machine learning. Cluster purity (PU) is a measure of the degree to which the induced role clusters meet the goal of containing only instances with the same GS role label:

$$Pu = \frac{1}{n}\sum_{i=1}^{n_C} \max_{j=1,\ldots,n_C} |C_i \cap C_j|$$

where C_i is an induced role cluster (a set of answers with the same semantic role label) and G_j is the best matching GS role cluster. Cluster collocation (CO) measures how well the clustering meets the goal of clustering all gold instances with the same label into a single predicted cluster:

$$Co = \frac{1}{n}\sum_{j=1}^{n_C} \max_{i=1,\ldots,n_C} |C_i \cap C_j|$$

The harmonic mean of PU and CO is reported as F-score (Lang, Lapata 2011; Fürstenau, Rambow 2012; Titov, Klementiev 2012).

---

[1] See Toldova et al. 2012 on morphological parsing and Gareyshina et al. 2012, Lyashevskaya et al. 2010 on dependency parsing.

In this article, we consider two other types of measures, taking into account (1) the distributional properties of semantic roles over the network of frames; (2) the path between two roles in a graph. From a common sense point of view, the pairs of roles like Instrument and Means, Theme and Patient are perceived as similar whereas Addressee and Reason, Experiencer and Direction are not. Meanwhile, it is important to distinguish the roles which can occur in the same frame such as Instrument and Means, Patient and Result. If the roles are distributed complementary over the frames of the same target verb, this allows us to downgrade the matching score between the NLP answer and GS.

Since our evaluation scheme is based on FrameBank (see next section), we will use the dictionary of valencies in order to calculate the co-occurrence statistics of the induced role RoleE and the corresponding gold standard RoleG. We compute the repulsion of roles as:

$$repulsion = \frac{\#Verbs(RoleE\_RoleG\ OR\ RoleE!RoleG)}{\sqrt{\#Verbs(Role\ E) \times \#Verbs(Role\ G)}}$$

Here, the numerator is the number of verbs in the dictionary for which the roles allow us to distinguish participants in the same frame (RoleE_RoleG, i.e. they co-occur in the same pattern: …V…RoleE…RoleG); plus the rest of verbs for which the roles distinguish frames within the same verb (RoleE!RoleG, i.e. the patterns like …V…RoleE… and …V…RoleG… are in complementary distribution). The repulsion is 0 if the roles do not compete with each other and 1 otherwise.

Thus, the roles Patient and Stimulus (of perception) can hardly stand in contrast to each other and their repulsion is expected to be low. However, we can easily anticipate a system in which both roles are labeled identically (e.g. as Patient). Quite the opposite, the pair Patient—Source_location appears to be a good candidate to have high repulsion since we expect to find lots of cases like *ja sorvalsja s dereva* 'I broke from the tree' where both roles co-occur in the same frame and label different kinds of participants (RoleE_RoleG). In addition, we can expect a number of cases such as *vino brodit* 'wine is fermenting'—*kochevniki brodili s mesta na mesto* 'the nomads wondered from place to place' where the roles Patient and Source_location are attested in different frames of the verb (RoleE!RoleG).

Furthermore, we assume that semantic roles are structured data which form a graph. If so, we can calculate the distance between the nodes just as the distances between senses in WordNet are calculated (cf. the review in Budanitsky, Hirst 2006). Presumably, these metrics will help us rank non-matching results as false alerts and major discrepancies (i.e. mistakes).

3. Data: Russian FrameBank

FrameBank is an open access database (www.framebank.ru) which consists of a dictionary of Russian lexical constructions (originally based on the valency dictionary Apresjan, Pall 1982) and a corpus of their uses tagged with a FrameNet-like annotation scheme (Lyashevskaya, Kuznetsova 2009, Lyashevskaya 2010, Kashkin,

Lyashevskaya 2013). FrameBank 1.0 offline release includes constructions for ca. 1500 frequent Russian verbs provided with up to 100 annotated examples per verb. Examples are randomly taken from the Russian National Corpus (RNC, http://ruscorpora.ru).

The theoretical framework adopted in FrameBank includes Construction Grammar (Ch. Fillmore, A. Goldberg, etc.) as well as some approaches developed in the Moscow Semantic School (Ju. D. Apresjan, E. V. Paducheva et al.). Another resource which has obviously influenced the Russian FrameBank is Berkeley FrameNet (http://framenet.icsi.berkeley.edu). However, in contrast to FrameNet, the core of FrameBank is constituted by the constructions of particular lexemes rather than by generalized frames. Each construction is stored in the dictionary as a pattern followed by a mnemonic sentence label. The pattern includes:

(1) the syntactic ranks and
(2) the morphosyntactic features of the arguments (incl. case and preposition marking),
(3) the semantic roles of the arguments,
(4) the lexical-semantic classes of the participants,
(5) the morphosyntactic features of the target lexical unit itself (e.g. impersonal, passive participle, etc),
(6) one or several examples.

Figure 1 shows a sample pattern in the dictionary.

The dictionary of constructions is supplemented by corpus examples tagged manually (see Fig. 2). An example is tied to a suitable pattern, which includes establishing correspondences between their elements, assigning morphosyntactic and semantic features of the arguments in a particular example, and also marking non-standard types of use (e.g., participial or converbal constructions). Adjuncts and focus particles are also tagged but remain beyond the construction pattern. The coordinates of phrases filling the slots and their heads are calculated automatically, so we can track the position of the filler against the predicate. If an expert comes across a corpus example which does not fit any existing pattern, they are expected to add a new pattern into the database.

There are two other components in FrameBank aimed at making generalizations on how the construction network of Russian verbs is organized. These are the graph of semantic roles and the graph of lexical constructions and frames. As regards the inventory of semantic roles, its volume and structure may shrink and expand depending on a particular research task and theoretical framework (see Fillmore 1968, 1977, 1982, Dowty 1991, Apresjan 1974/1995: 125–126, Apresjan et al. 2010: 370–377, Paducheva 2004: 587–588, etc.). The most important principles governing the inventory of semantic roles in FrameBank are as follows (Kashkin, Lyashevskaya 2013):

- the inventory should be hierarchical in order to support flexible search options (it may be reduced to 5–10 basic roles, and at the same time enlarged to several dozen labels);
- the roles should correlate with the semantic classification of verbs (what follows from it is that traditionally "broad" roles such as Agent or Patient should get different labels in different semantic classes, cf. Agent in destruction vs. speech vs. motion).

	Synt_Rank	Morph.	Semantic role	Lexical class
svesti	Predicate	Vimpers		
Y	Object	Sacc	Part of subject of physiological state	body part
X	Periphery	ot + Sgen	Reason	abstract

Fig. 1. The pattern of the construction *Pal'tsy*$_Y$ *svelo*$_V$ *ot xoloda*$_V$ 'The fingers$_Y$ cramped$_V$ from the cold$_X$'

	Synt_Rank	Morph.	Semantic role	Lexical class	Alternation predicted by	Filler
X	Subject	Snom	Reason	abstract		
	Periphery	Sins	=	=	Passive participle	*prostudoj*
Svesti	Predicate	V				
	Predicate. attrib	V.partcp. pass.full. acc				*svedennyje*
Y	Object	Sacc	Part of subject of physiological state	body part		
	Agreement controller	Sacc	=	=	Passive participle shift	*pal'tsy*

Fig. 2. The annotation of the construction *Sudoroga*$_X$ *svela*$_V$ *pal'tsy*$_Y$ 'A cramp$_X$ in (lit. took down$_V$) the fingers$_Y$' in the example ... *ona podsela k pechi, svedennyje*$_V$ *prostudoj*$_X$ *pal'tsy*$_Y$ *zasovyvala v samyj ogon'—grela* '... she sat down next to the stove trying to warm at the fire her fingers$_X$ cramped$_V$ by flu$_X$'. For each element, the first line reports data from the dictionary, the second line reports annotation of the example.

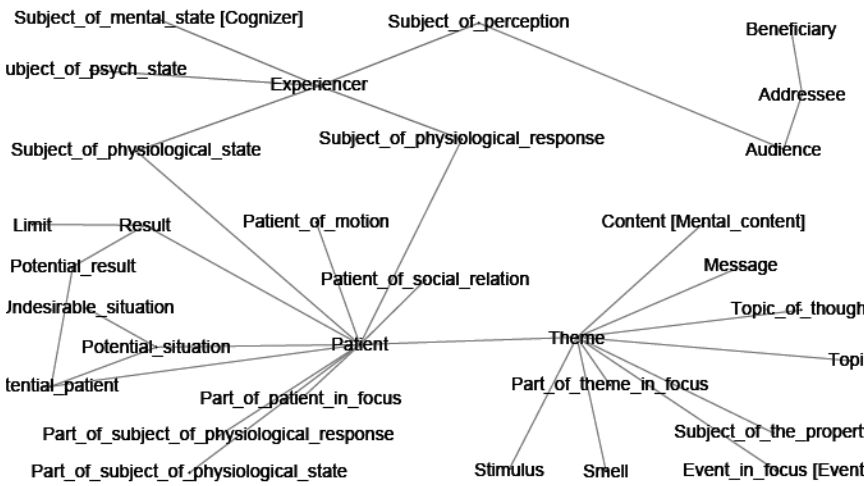


Fig. 3. A fragment of the sematic roles graph illustrating the domains of Experiencer, Addressee and Patient

The detailed list of semantic roles currently contains 96 items classified into six domains (those of Agent, Patient, Experiencer, Instrument, Addressee, Settings), which are further subdivided into smaller units. As an instance, the domain of Experiencer includes Subject of Perception ('see', 'hear'), Subject of Mental State ('think', 'understand'), Subject of Psychological State ('love', 'be afraid'), Subject of Physiological State ('feel pain', 'have a buzzing in one's ears'), and Subject of Physiological Response ('laugh', 'feel sick'). In addition, the last two roles are linked to the node of Patient whereas the subjects of Perception, Mental State and Psychological State are linked to the node of Agent. As a result, all the roles make up a united graph, see Figure 3.

The graph of lexical constructions and frames is an ongoing project; for the present it has covered 55% of the data. The graph of constructions documents the systematic relations between constructions. First, it systematizes semantic shifts in verbal lexemes (metaphor, metonymy and some more complex relations). Second, the graph represents formal changes in argument structure, such as omission of a participant, diathetic alternations, the inheritance of a pattern from another verb etc. The semantic part of the project is inspired by FrameNet grapher as well as by E. Rakhilina et al.'s research on Russian polysemous adjectives and adverbs summarized in a database (see Rakhilina et al. 2010 and references therein). The formal part is guided by E. Paducheva and G. Kustova's theoretical and empirical analysis of polysemy in Russian verbs (Paducheva 2004, see also the Lexicographer database at http://lexicograph.ruslang.ru) The frame grapher shows how lexical constructions map into the frame structure, so the graph of lexical constructions goes in parallel with the graph of frames.

4. Experiment

The next two sections will focus on how the FrameBank data can be used in SRL evaluation. The goal of our experiment was to build up a simplified SRL system adapted to use FrameBank data as a source and to do a basic evaluation in both quantitative and qualitative terms.

4.1. SRL Prototype

In order to get a data set for the evaluation we produced a list of 62 heuristics simulating a rule-based SRL tool. Unlike machine learning tools adapted to corpus data, our system works with data from the FrameBank dictionary of constructions. Our experiment focuses on semantic role labeling of four prepositional phrases: *za* + NPins, *za* + NPacc, *ot* + NPgen, *po* + NPdat. These particular PPs have been chosen since they are very frequent (e.g. ca. 900,000 hits of the PP *ot* + NPgen in the RNC) and highly polysemous. FrameBank annotations show that za + NPacc is mapped into 14 roles, such as Destination Point (*Mal'chik zabezhal za derevo* 'The boy ran behind the tree'), Motivation (*nakazat' syna za vran'je* 'to punish a son for his lies'), Price (*On kupil dom za million dollarov* 'He bought a house for a million dollars'), Period (*Eto mozhno sdelat' za chas* 'It can be done in an hour'), etc.

We produced a list of heuristics that take into account the morphosyntactic pattern of the construction, the lexical class of the PP argument, the lexical class of other arguments and the lexical class of the target predicate (the similar feature types were used for the rule-based verb sense disambiguation on RNC data, see Toldova et al. 2008).

The rules can be illustrated by the following two uses of *za* + NPacc. If the PP is added to a transitive construction, the target verb refers to the change of a possessor (e.g. *kupit'* 'to buy', *prodat'* 'to sell', *otdat'* 'to give', etc.) and the NPacc embedded into *za* + NPacc is a quantitative expression (e. g. *sto rublej* 'one hundred rubles', *bol'shaja summa* 'a large sum [of money]'), then the semantic role of *za* + NPacc is Price. However, if the class of the NPacc (within za + NPacc) is a time period (*dva dnja* 'two days', *nedelja* 'a week', etc.), then the semantic role of the PP in focus is Period.

Some rules suggest two possible outcomes, cf. two constructions: *Militsioner pobezhal za prestupnikom* 'A policeman ran after an offender', where the pragmatically correct choice is Counter-Agent, and *Mal'chik pobezhal v bol'nitsu za vrachom* 'A boy ran to hospital to call the doctor', where *za* + NPacc is more likely to describe Goal, because the doctor does not seem to escape or perform any other action here. This distinction is hard to formalize, as it requires taking into account rather vague pragmatic factors, so the rule assigns two roles with a 50% probability when the NPacc in the PP is animate.

4.2. Training and Test Data

The rules were formulated for the constructions found in the so-called 'old' part of the dictionary and evaluated against the 'new' constructions. As a training set, the constructions attested in Apresjan, Pall 1982 were taken. The constructions recently added by annotators (in order to cover RNC examples) were used as a test data set. Though this was but one of many possible folds, yet the split between the 'old' and 'new' parts was chosen as a matter of convenience since we presumed the productive patterns to prevail in the new part. Table 1 shows the distribution of training and test data set[2].

[2] Patterns with generalized locative and directional PPs were also taken into account if the use of one of the four PPs was attested in FrameBank. E.g. such patterns as [NPnom] V [za + NPacc] include those of [NPnom] V [PRkuda + NPx], where PRkuda stands for any directional preposition corresponding to Russian kuda 'where (direction)', and x denotes the NP case governed by a particular preposition.

Table 1. Type frequency (the number of constructions in the dictionary) and token frequency (the number of annotated sentences in corpus samples) of four Russian PPs: training and test set

	Training set: 'old' data		Test set: 'new' data	
PP	**# constructions**	**# examples**[3]	**# constructions**	**# examples**
za + NPins	95	80	19	22
za + NPacc	228	223	37	51
ot + NPgen	266	435	70	113
po + NPdat	311	245	65	78
Total	900	983	191	264

The results of SRL for four prepositional groups based on our rules are shown in Table 2.

Table 2. Results of the experiment

PP	Total amount of new patterns	'Strong' matching (the role is identified correctly and unambiguously	'Weak' matching (one of the answers is correct)	P_{strong}	$P_{strong+weak}$
za + NPins	19	9	7	0.47	0.84
za + NPacc	37	22	11	0.59	0.89
ot + NPgen	70	41	24	0.59	0.93
po + NPdat	65	32	25	0.49	0.88
Total	191	104	67	0.54	0.90

4.3. SRL Cues

The rules for disambiguation produce the right answers due to taking into account such 'cues' in the data as:

- The semantic class of a verb. Thus, the semantic role of *za* + NPacc has been identified correctly as Reason for an Emotion in the pattern *Beshus' za doch moju* 'I am in a rage because of my daughter', because the verb *besit'sja* 'to be in a rage' used here belongs to the class of emotions, like the verbs *bespokoit'sja* 'to worry about sth.', *bojat'sja* 'to be afraid' etc. which also occur in this syntactic pattern. Similarly, *ot* + NPgen is interpreted as Reason when combined with verbs denoting

[3] The quantity of annotated corpus examples can be less than the quantity of constructions in the dictionary since FrameBank is a project in progress and not all constructions has been tagged so far. For this reason, the rules and evaluation are based on types (i.e. constructions in the dictionary), not tokens.

physiological state, cf. the new construction patterns *V golove gudelo ot udara* 'One was feeling a buzzing in one's head due to the stroke' or *Vo rtu gorelo ot pertsa* 'One's mouth was burning from the pepper' and the old ones *Ushi zalozhilo ot vystrelov* 'One's ears were blocked due to the shots' or *Zhivot podvelo ot goloda* 'One was feeling pinched with hunger (lit.: It brought one's stomach closer due to hunger').

- The semantic class of a participant. For example, in the case of *Po radio igrala muzyka* 'There was music broadcast (lit.: played) by radio' the role of Manner has been assigned to *po* + NPdat, since NPdat in this case is in the semantic class of Communication Facilities, cf. *zvonit' po telefonu* 'to ring sb. up', *vystupat' po televizoru* 'to speak on TV', *poslat' dokumenty po pochte* 'to send documents by post', etc. The same idea works in numerous cases where a participant must be animate (like Agent or Counter-Agent), as well as in the case of the opposition between concrete and abstract entities which is relevant for quite a few examples.
- The pattern in general. Sometimes it is necessary to take into account the interaction between the elements of a construction. A curious example is represented by the verb *begat'* 'to run' and its prefixal derivative *probegát'* when used with *po* + NPdat. Normally they refer to motion events in a syntactic pattern NPnom V *po* + NPdat (*Rebenok begaet po komnate* 'A child is running in the room'). Used metaphorically, these verbs may describe perception, which may be supported by adding a NPins *glazami* 'with one's eyes' or *vzgljadom* 'with one's look' into the pattern, cf. *Ona probezhala glazami po tekstu pis'ma* 'She looked through (lit.: ran with her eyes) the text of the letter' or *Ona probezhala glazami po komnate* 'She looked through (lit.: ran with her eyes) the room'—note that in the latter case *po* + NPdat refers to the same entity (the room) as it does in the situations of motion. What is the main point here is that it is possible to omit a NPins in the contexts of perception, but a NPdat embedded into *po* + NPdat cannot denote then any kind of territory or space. Thus, *Ona probezhala po tekstu pis'ma* 'She looked through (lit.: ran) the text of the letter' is perfect, while *Ona probezhala po komnate* 'She ran along the room' is very odd as a reference to visual perception.

4.4. Challenges for SRL

The main challenges we have faced in our experiment are as follows.

First, it is difficult to deal with such cases in which there are no clear constraints on the classes of a verb and of its arguments. Thus, the use of *po* + NPdat for conveying Reason in *Rasskaz byl prochitan po ego pros'be* 'The story was read at his request' receives an additional interpretation of Information Source (yielded by the semantics of the noun *pros'ba* 'request', which belongs to the class of texts and speech acts, and of the verb *prochitat'* 'to read' dealing with information processing). This case of ambiguity stems from the vagueness of semantic restrictions imposed on *po* + NPdat as Reason, as well as on the verbs possible in this pattern, cf. *zhenit'sja po ljubvi* 'to make a love-match (lit.: to get married due to love)', *uvolit' po sokrashcheniju shtatov* 'to discharge sb. on grounds of staff reduction', *sidet' zdes' po drugomu delu* 'to stay here on some other business'.

Second, a challenge is posed by metonymic shifts of concrete nouns. For example, *ot* + NPgen in a new pattern *Ego nevozmozhno otorvat' ot knigi* 'It is impossible to divert his attention from the book (lit.: to tear him=it from the book)' is wrongly analyzed not as Content of Action, but as Patient & Location (like in *otorvat' listok ot kalendarja* 'to tear a sheet off the calendar'). The role of Content of Action presumes an abstract entity (*otorvat' ot raboty* 'to put sb. off work', *otkazat'sja ot svoih planov* 'to abandon one's plans'), while in the example in question this kind of participant is referred to by a concrete noun *kniga* 'a book' metonymically connected with an abstract entity of reading which is meant here.

There are some more similar examples in the training corpus which pose an obvious difficulty for constructing the rules. For instance, the verb *sidet'* 'to sit' combined with *za* + NPins may refer to sb's posture when the NPins denotes a concrete entity, and *za* + NPins takes the role of Location (*Papa sidit za knizhnym shkafom* 'Father is sitting behind the bookcase'), or it may yield the interpretation of *za* + NPins as Content of Action if the NPins designates an abstract entity (*Papa sidit za rabotoj* 'Father is occupied with his work (lit.: Father is sitting behind his work)'). The latter interpretation may however arise in the case of a concrete NPins making it difficult to automatically distinguish it from locative contexts, cf. the sentence *Papa sidit za knigoj* (lit.: 'Father is sitting behind the book'), which means that father is occupied with reading (as a result an obvious metonymic connection between reading and books) and has nothing to do with the expression of a spatial relationship between father and the book. Even a more challenging example is *Papa sidit za stolom* 'Father is sitting at the table', which evokes a dual interpretation (Location vs. Content of Action—e.g., eating) depending perhaps on a broader context.

It can be seen from the above that what influences the choice of a semantic role is pragmatic factors, which have proved to cause difficulties for the application of our rules. This can be illustrated by the use of *za* + NPins in the roles of Counter-Agent vs. Goal in the frames of motion when the NPacc participant is animate. Thus, a new pattern *Otets pustilsja za gigantskoj akuloj v nebol'shoj motornoj lodke* 'Father rushed after a giant shark in a small motorboat' gets two interpretations, Counter-Agent and Goal.

5. Evaluation at Work. Discussion

The experimental rule-based SRL module for four PPs yields P=0.90 in trade-off evaluation (a rule can induce more than one role as an answer, one of them is correct) and P=0.54 in strong evaluation (exact matching of an answer with the GS); recall is not applicable because we used default settings in our rules. We did not apply the formulae of purity and collocation since there was a very small number of data points to perform cluster modeling.

Table 3 summarizes 15 cases of non-matching answers. The non-matching pairs of roles were tagged manually as Good, Average and Bad match by an assessor (the contexts are provided in the table). For each case, we show the statistics on the co-occurrence of semantic roles in the valency dictionary; the shortest path from RoleE to RoleG in the graph of semantic roles. The borders of domains are marked by square brackets, ↑ marks the path up to the hypernym and ↓ marks the path down to the hyponym. The

last column shows whether the roles belong to the same domain (e.g. Agent, Patient, etc); label (Yes) indicates that the roles are from the same hyper-domain of settings (circumstances) but they belong to different domains such as Place, Time, Reason etc.

Table 3. The goodness of fit for non-matching pairs of roles: manual evaluation and descriptive statistics

Matching Evaluation (human)	Role E	Role G	#Verbs (RoleE)	#Verbs (RoleG)	#Verbs (RoleE! RoleG)	#Verbs (RoleE+ RoleG)	Repul-sion	Same domain
Good	Source	Reason	12	266	3	0	0.05	NO
[Source↑External_cause↑Agent]↑Root↓[Setting↓Reason] *Lovit' kajf* [*ot knig*] 'To be in high from books'.								
Good	Path	Patient	105	712	46	3	0.18	NO
[Path↑Location↑Setting]↑Root↓[Patient] *On bredit i mechetsja golovoj* [*po perekladine*] 'He raves, tossing his head over the crossbar'.								
Good	Property	Reason	175	266	31	5	0.17	(YES)
[Property↑Setting↓Reason] *Ego zabrali* [*po natsional'nomu piznaku*] 'He was arrested on ethnic grounds'.								
Average	Term	Time_ point	52	42	6	2	0.17	YES
[Term↓Time_point] *Vstrecha prodlilas'* [*za polnoch*] 'The meeting lasted past midnight'.								
Average	Term	Target_ location	52	398	26	0	0.18	(YES)
[Term↑Time↑Setting↓Location↓Target_location] *Emu zabralos'* [*za 50 let*] 'He was (lit. It was climbed him) over fifty years old'.								
Average	Source_ of_smell	Source_ location	5	250	2	0	0.06	NO
[Source_of_smell↑Source]↓[Resourse↑Sourse_location] [*Ot tebja*] *za verstu paxnet neprijatiem sotsialisticheskix tsennostej* 'It is evident from a mile away that you reject (lit. it smells from you) socialist values'.								
Average	Source_ location	Poten-tial_coun-ter-agent	250	6	1	0	0.03	NO
[Sourse_location↑Location↑Setting]↑Root↓[Agent↓Counter-agent↓Potential_counter-agent] *Devochki glupo prygali* [*ot nego*] *v trolleybus* 'The girls foolishly jumped from him on a trolleybus'.								
Average	Undesir-able_sit-uation	Event_ in_focus	32	309	7	0	0.07	YES
[Undesirable_situation↑Potential_situation↑Result↑Patient↓Theme↓Event_in_focus] *On uderzhalsja* [*ot sljoz*] 'He hold back the tears'.								
Bad	Patient	Source_ location	712	250	146	127	0.65	NO
[Patient]↑Root↓[Setting↓Location↓Sourse_location] *Ona otorvala glaza* [*ot knigi*] 'She raised her eyes from the book'.								

Matching Evaluation (human)	Role E	Role G	#Verbs (RoleE)	#Verbs (RoleG)	#Verbs (RoleE! RoleG)	#Verbs (RoleE+ RoleG)	Repul-sion	Same domain
Bad	Patient	Manner	712	320	175	91	0.56	NO
[Patient]↑Root↓[Instrument↓Manner] *Probejte* [*po baze dannyh*] *ego prava* 'Check his license status through the database'.								
Bad	Location	Counter-agent	519	285	117	12	0.34	NO
[Location↑Setting]↑Root↓[Agent↓Counter-agent] *Povtorit'* [*za uchitelem*] 'To repeat after the teacher'.								
Bad	Purpose	Location	169	519	84	2	0.29	NO
[Purpose↑Setting↓Location] *Deti prygali by* [*na mogile*] 'His kids would jump on the grave'.								
Bad	Informa-tion_re-source	Message	26	236	12	10	0.28	NO
[Information_resource↑Resourse↑Sourse_location↑ Location↑Setting]↑Root↓[Patient↓Theme↓Message] *Obychaj zvat' doma* [*po familii*] 'The tradition to address by the last name at home'.								
Bad	Informa-tion_re-source	Manner	26	320	9	0	0.10	NO
[Information_resource↑Resourse↑Sourse_location↑Location↑Setting]↑Root↓[Instrument↓Manner] *Izbirateli golosujut* [*po spiskam*] 'The voters take a vote through the lists'.								
Bad	Informa-tion_re-source	Cause	26	144	7	0	0.11	NO
[Information_resource↑Resourse]↑[Source↓External_cause↓Agent↓Cause] *Ja prochitala pis'mo* [*po ego pros'be*] 'I read the letter at his request'.								

The figures of repulsion correlate quite well with the split between Average and Bad matches (repulsion threshold at about .20). However, the formula fails to predict the split between Good and Average matches.

Even though all the Bad matches do not share the same domain, this factor was not sufficient to assess the quality of mismatches. We experimented with a number of approaches to employ a graph-based distance model, but neither of them succeeded to rank the data in accordance with manual assessment.

Nevertheless, the use of graph seems to be promising for the task of 'smart' SRL evaluation. In fact, the nature of mismatches in our experimental set is not necessarily the same as expected with the real SRL data since the chance of deviation from the training model is much higher. Token observations will provide a much larger number of data points and more homogeneous results.

Still, there is a lot of work to be done to explore the answers of external parsers with non-matching grids of semantic roles. *Further goal is to integrate a better representation of* the graph of semantic roles. So far, the graph uses two types of edges (IS-A and association). As a rule, the hyponyms cannot co-occur with their hypernym in the same frame (cf. Agent, Cause, Speaker), so another type of edges would be useful.

References

1. *Apresjan Ju. D.* (1995), Selected papers, Vol. 1, Lexical Semantics [Izbrannye trudy, tom I. Leksicheskaja semantika], Jazyki Russkoj Kul'tury, Vostochnaja Literatura, Moscow.
2. *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.
3. *Apresjan Ju. D., Pall E.* (1982), Russian verb—Hungarian verb. Government and combinability [Russkij glagol—vengerskij glagol. Upravlenie i sochetaemost'], Tankyonvkiado, Budapest.
4. *Azarowa, I.* (2008). RussNet as a computer lexicon for Russian. Intelligent Information Systems, pp. 341–350.
5. *Baker C., Ellsworth M., Erk K.* (2007) SemEval'07 task 19: Frame semantic structure extraction, Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, June 2007, pp. 99–104.
6. *Budanitsky A., Hirst G.* (2006), Evaluating WordNet-based measures of lexical semantic relatedness, Computational Linguistics, 32 (1), pp. 13–47.
7. *Das D., Chen D., Martins A. F. T., Schneider N., Smith N.* (forthcoming), Frame-semantic parsing. Computational Linguistics.
8. *Dowty, D. R.* (1991), Thematic proto roles and argument selection, Language 67, pp. 547–619.
9. *Ermakov A. E., Pleshko V. V.* (2009), Semantic interpretation in text processing computer systems [Semanticheskaja interpretatsija v sistemah komp'juternogo analiza teksta]. Information technologies [Informatsionnye tehnologii], Vol. 6, pp. 2–7.
10. *Fillmore Ch. J.* (1968), The Case for Case, in Bach E. and Harms (Ed.), Universals in Linguistic Theory. New York, pp. 1–88.
11. *Fillmore Ch. J.* (1977), The case for case reopened, in Cole P., Sadock J. M. (eds.), Grammatical Relations, Acad. Press, New York, pp. 59–81.
12. *Fillmore Ch. J.* (1982), Frame semantics, Linguistics in the morning calm: Selected papers from the SICOL-1981, Hanship, Seoul, pp. 111–137.
13. *Fillmore, Ch. J., Johnson C. R., Petruck M. R. L.* (2003), Background to FrameNet. International Journal of Lexicography, 16(3), pp. 235–250.
14. FrameNet. An online resource, available at: http://framenet.icsi.berkeley.edu.
15. *Fürstenau H., Rambow O.* (2012), Unsupervised induction of a syntax-semantics lexicon using iterative refinement, in Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM 2012).
16. *Gareyshina, Anastasia, Maxim Ionov, Olga Lyashevskaya, Dmitry Privoznov, Elena Sokolova, Svetlana Toldova.* (2012). RU-EVAL-2012: Evaluating dependency parsers for Russian. In: Proceedings of COLING 2012, Mumbai, December 2012: Posters. Pp. 349–360.
17. *Gorinski, P., Ruppenhofer, J., Sporleder, C.* (2013), Towards weakly supervised resolution of null instantiations. In Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013). Long Papers, pp. 119–130. http://aclweb.org/anthology//W/W13/W13-0111.pdf

18. *Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M. A., Màrquez, L.* et al. (2009), The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task, pp. 1–18. http://aclweb.org/anthology//W/W09/W09-12.pdf
19. *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.
20. *Kuznetsov I.* (2013), Semantic role labeling system for Russian language, in: Joho H., Ignatov D. (eds.), ECIR 2013 Doctoral Consortium, 24 March 2013, Moscow, pp. 15–18. http://www.hse.ru/pubs/lib/data/access/ticket/1391119334ea59778f895f58081f821387ee54322f/text_DC.pdf
21. *Lang J., Lapata M.* (2011), Unsupervised semantic role induction with graph partitioning, Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1320–1331.
22. *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.
23. *Lyashevskaya O., Astaf'eva I., Bonch-Osmolovskaya A.* et al. (2010), NLP evaluation: Russian morphological parsers [Otsenka metodov avtomaticheskogo analiza teksta: morfologicheskije parsery russkogo jazyka], Computational linguistics and intellectual technologies. Proceedings of International Conference "Dialog", Vol. 9 (16), pp. 318–326.
24. *Lyashevskaya O. N., Kuznetsova, Ju. L.* (2009), Russian FrameNet: constructing a corpus-based dictionary of constructions [Russkij FrejmNet: k zadache sozdanija korpusnogo slovarja konstruktsij], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog", Vol. 8 (15), pp. 306–312.
25. *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.
26. *Paducheva E. V.* (2004), Dynamic patterns in lexical semantics [Dinamicheskie modeli v semantike leksiki], Jazyki slavjanskoj kul'tury, Moscow.
27. *Palmer, M. S., Gildea D., Kingsbury P.* (2005), The proposition bank: an annotated corpus of semantic roles. Computational Linguistics, 31(1), pp. 71–106. www.cs.rochester.edu/~gildea/palmer-propbank-cl.pdf.
28. *Palmer M. S., Wu Sh., Titov I.* (2013), Semantic Role Labeling Tutorial. NAACL 2013 tutorials. Electronic access: 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
29. *Petrova M.* (2013). The Compreno Semantic Model: The Universality Problem. International Journal of Lexicography, 26 (4).

30. *Rakhilina E. V., Reznikova T. I., Karpova O. S.* (2010), Semantic shifts in attributive constructions: metaphor, metonymy, and rebranding [Semanticheskie perehody v atribultivnyh konstruktsijah: metafora, metonimija i rebrending], in Linguistics of Constructions [Lingvistika konstrukcij], Azbukovnik, Moscow, pp. 398–455.
31. *Smirnov I. V., Shelmanov A. O., Kuznetsova E. S. and Hramoin I. V.* (2014), Semantic-syntactic analysis of natural languages. Part II. Method for semantic-syntactic analysis of texts [Semantiko-sintaksicheskij analiz estestvennyh jazykov II. Metod semantiko-sintaksicheskogo analiza tekstov]. Artificial Intelligence and Decision Making [Iskusstvennyj intellekt i prinjatie reshenij], Vol. 1, pp. 95–108.
32. *Surdeanu M., Johansson R., Meyers A., Màrquez L., Nivre J.* (2008), The CoNLL 2008 shared task on joint parsing of syntactic and semantic dependencies, Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL-2008). Manchester, England, August 2008, pp. 159–177. http://aclweb.org/anthology//W/W08/W08-2121.pdf
33. *Titov I., Klementiev A.* (2012), Semi-supervised semantic role labeling: approaching from an unsupervised perspective, in Proceedings of COLING 2012: Technical Papers, pp. 2635–2652.
34. *Toldova S. Ju., Kustova G. I., Lashevskaja O. N.* (2008), Semanticheskie fil'try dlja razreshenija mnogoznachnosti v Natsional'nom korpuse russkogo jazyka: glagoly [Semantic filters for word sense disambiguation in the Russian National Corpus: verbs], Computational linguistics and intellectual technologies. Proceedings of International Conference “Dialog”. Vol. 7 (14), pp. 522–529.
35. *Toldova S. Ju., Sokolova E. G., Astaf'eva I., Gareyshina A., Koroleva A., Privoznov D., Sidorova E., Tupikina L., Lyashevskaya O. N.* (2012). NLP evaluation 2011–2012: Russian syntactic parsers [Otsenka metodov avtomaticheskogo analiza teksta 2011–2012: Sintaksicheskie parsery russkogo jazyka]. Computational linguistics and intellectual technologies. Proceedings of International Conference “Dialog”, Vol. 11 (18), pp. 797–809.

ДАННЫЕ ИНТЕРНЕТА В ИССЛЕДОВАНИИ ЯЗЫКОВЫХ ИЗМЕНЕНИЙ: АНАЛИЗ ЧЕРЕДОВАНИЙ В РУССКИХ КОМПАРАТИВАХ И ПРОГРАММА ДЛЯ РАБОТЫ С ТАКИМИ ДАННЫМИ

Магомедова В. Д. (varya.magomedova@gmail.com)
Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Слюсарь Н. А. (slioussar@gmail.com)
Национальный исследовательский университет Высшая школа экономики, Москва, Россия
Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Ключевые слова: компаратив, согласные, исторические чередования, оптимизация поиска

INTERNET DATA IN THE STUDY OF LANGUAGE CHANGE: A CASE STUDY OF ALTERNATIONS IN RUSSIAN COMPARATIVES AND A PROGRAM TO WORK WITH SUCH DATA

Magomedova V. D. (varya.magomedova@gmail.com)
Saint Petersburg State University, St. Petersburg, Russia

Slioussar N. A. (slioussar@gmail.com)
National Research University
Higher School of Economics, Moscow, Russia
Saint Petersburg State University, St. Petersburg, Russia

The Internet is a unique source of non-standard forms, which gives us a novel opportunity to analyze fine-grained dynamics of language change. We used this opportunity to study the decay of historic consonant

alternations in Russian. In standard Russian, these alternations are present in some verb forms and in comparatives (e.g. *suxoj* 'dry' — *sushe* 'drier', *ljubit'* 'to love' — *ljublju* 'I love'), as well as before certain derivational suffixes. Verb forms have been recently studied by Slioussar and Kholodilova (2013), and we looked at comparatives. Two groups of adjectives were selected: ones that have normative comparatives with alternations and ones that do not, but native speakers still try to generate such forms. In the first group, some adjectives like *ubogij* 'poky' have up to 30 % of comparatives without alternations, but, unlike with verbs, no significant correlation with adjective frequency or its other characteristics was found. The second group consisted primarily of compound adjectives ending in *-gij, -kij, -xij*. Here, the most important factor is whether the second part of the compound is used as an independent adjective. If it is not (e.g. as in *dlinnorukij* 'long-armed'), most comparatives lack alternations. Searching for forms on the Internet, we faced many problems. The counts provided by search engines are extremely inaccurate, only the first thousand results are shown, they cannot be downloaded in a convenient format, contain a lot of typos and other irrelevant data etc. We present a program called *Lingui-Pingui* that we developed to solve these and some other problems.

Keywords: comparative, consonants, historical alternations, search optimization

1. Введение

Огромное количество нередактируемых текстов в интернете дает лингвистам новые возможности для наблюдения за жизнью языка. С приходом массовой грамотности процессы языковых изменений значительно замедлились, и всё-таки удается найти интересные примеры таких явлений. В этой работе мы анализируем процесс разрушения исторических чередований согласных, система которых остается неизменной в нормативном русском языке[1], но явно расшатывается в целом ряде ненормативных форм. Мы рассматриваем это явление на примере форм компаратива.

Однако при использовании данных из интернета возникает целый ряд проблем, связанных с их сбором и обработкой. Например, статистика по результатам поиска, которую дают поисковые машины, очень неточная, что существенно осложняет оценку частотности различных форм. При этом просматривание результатов вручную затрудняется тем, что их невозможно сохранить в удобной для работы форме. Мы разработали программу «Lingui-Pingui», которая работает с Yandex.API и помогает собирать, сортировать и обрабатывать данные, полученные из Интернета. Цель данной работы двояка: представить результаты анализа форм компаратива и созданную нами программу «Lingui-Pingui».

---

[1] В нормативном русском языке исторические чередования согласных (*к // ч, к // ц, г // ж, г // з* и др.) сохраняются в ряде глагольных классов (например, *писать — пишу, возить — вожу*), в некоторых формах компаратива (например, *сухой — суше*), а также перед определенными суффиксами (например, *скользить — скольжение*).

2. Предыдущие исследования

Н. А. Слюсарь и М. А. Холодилова [Slioussar, Kholodilova 2013] изучали исторические чередования согласных на материале глаголов. Они провели анализ ненормативных форм, которые можно в изобилии найти в интернете, и выяснили, что система чередований расшатывается. Это явление особенно интересно, так как парадигматическое выравнивание давно и активно обсуждается в рамках самых разных лингвистических направлений [например, Albright 2002, 2010; Anttila 1977; Benua 1997; Bybee 1985; Kiparsky 1982, 2002; Kuryłowicz 1949; Mańczak 1958; McCarthy 2005].

Слюсарь и Холодилова показали, что процесс парадигматического выравнивания зачастую идет в двух направлениях одновременно: избавление от чередований в тех формах, где они были, и появление чередований в тех, где их не было. Однако избавление от чередований более распространено. В этой работе также были выведены факторы, влияющие на сохранность чередований, на примере различных глаголов единственного продуктивного класса, где встречаются чередования, — 1 подкласса X класса. Это частотность, нормативность глагола (у редких и ненормативных глаголов больше проблем с чередованиями) и конечный согласный основы. Чередования сохраняются лучше всего у глаголов с основами на губные согласные и хуже всего у глаголов с основами, заканчивающимися группами шумных согласных. Кроме того, были найдены примеры чередований, не встречающихся в нормативном русском языке. В этом случае мы имеем дело с интересной ситуацией, когда носитель языка понимает, что в той или иной форме необходимо чередование, но уже не уверен, какое именно.

3. Интересующие нас формы компаратива и различные их свойства

1. Чередования согласных встречаются только в неизменяемых синтетических формах компаратива с суффиксом *-е*. Другие суффиксы, при помощи которых образуются эти формы, *-ее/ей* и *-ше,* чередований не требуют. Суффикс *-е* присоединяется ко всем прилагательным с основой на *-г, -к, -х*, у которых есть синтетические формы, и к нескольким основам на *-д, -т* (кроме того, есть единичные случаи вроде *дешевле*). При этом у некоторых прилагательных на *-кий* выпадает суффикс *-(о)к-*.
2. В данной работе мы рассмотрим два типа ненормативных форм: от прилагательных, у которых есть нормативный компаратив с чередованием, и от прилагательных, у которых его нет. Отсутствие синтетического компаратива может быть обусловлено как семантикой прилагательного (в этом случае не образуются никакие формы компаратива), так и его морфологическими характеристиками. Более подробную информацию об этом можно найти, например, в работах [Зализняк 1977; Князев 2007; Русская грамматика 1980 и др.]. Судя по нашим наблюдениям, в тех случаях, когда

образование компаратива допустимо с точки зрения семантики, носители русского языка имеют тенденцию употреблять не только аналитические, но и ненормативные синтетические формы. Следует заметить, что небольшая часть текстов, где встречаются такие формы, имеет шутливый характер. Так как однозначно отделить такие тексты от прочих не представлялось возможным (не всегда понятно, шутит ли пишущий), мы учитывали их вместе со всеми. В целом у нас создалось ощущение, что люди склонны чаще шутить с теми же формами, в которых чаще встречаются ошибки.

3. Наша задача заключалась в том, чтобы проанализировать распределение форм с чередованиями и без них от различных прилагательных. Мы искали как минимум десять форм для каждого прилагательного, проиллюстрированных ниже на примере прилагательного *упругий*: с чередованиями и без них, с суффиксами *-е* и *-ее/ей*, с приставкой *по-* и без нее. В принципе, утеря чередований сопровождается заменой суффикса *-е* на более новый продуктивный *-ей/ее*, но мы заметили, что эти процессы не всегда идут параллельно и включили промежуточные формы с чередованиями с суффиксом *-ее/ей*. Затем наше внимание привлек тот факт, что от некоторых прилагательных встречаются и формы без чередований с суффиксом *-е*, например, *круте* или *сухе*. Мы уже не успели включить их в основное исследование, поэтому в разделе 4 только перечисляем те (немногочисленные) прилагательные, от которых были найдены такие формы. Для прилагательных с суффиксами мы также искали формы, где эти суффиксы выпадают, например, *хлипее* наравне с *хлипкее*.

(1) *_упруже, поупруже, упругее, упругей, поупругее, поупругей, упружее, упружей, поупружее, поупружей*

Важно заметить, что проблемы с чередованиями могут проявляться и в избегании синтетических форм за счет использования аналитических. Мы не анализируем соответствующие данные в этой статье, но планируем сделать это в будущем.

4. Распределение форм компаратива с чередованиями и без них

Согласно «Грамматическому словарю русского языка» [Зализняк 1977], в русском языке 113 бесприставочных прилагательных с основами на *-г, -к, -х*, образующих нормативные формы компаратива с чередованием. Ненормативные формы от этих прилагательных довольно редки[2], и результаты поиска этих форм, как правило, содержат много «мусора» — опечаток и других нерелевантных данных. Поэтому мы отобрали 23 прилагательных, у которых результаты

[2] Как показывают в своей работе Н. А. Слюсарь и М. А. Холодилова [Slioussar, Kholodilova in 2013], похожая ситуация наблюдается и с глаголами: ненормативные формы от глаголов, относящихся к литературному русскому языку, также редки.

первичного поиска нестандартных форм (одна форма без чередования на *-ее* для каждого прилагательного) содержали не менее 4 % релевантных данных. К этим 23 прилагательным были добавлены девять с основами на *-т, -д* (все бесприставочные с чередованиями кроме *худой* в значении «плохой»). Также мы рассматривали прилагательные, которые не имеют нормативных синтетических компаративов: 13 сложных прилагательных с основами на *-г, -к, -х* (например, *близорукий*), несколько слов с суффиксами *-ск-* и *-цк-* и прилагательное *великий*.

Распределение различных форм у прилагательных первой и второй группы представлено ниже в Таблицах 1–4. Данные о частотности прилагательных взяты из «Частотного словаря современного русского языка» [Ляшевская, Шаров 2009], у тех прилагательных, которые не вошли в этот словарь, в соответствующей графе стоит прочерк. Количество результатов указывается после фильтрации.

Формы с суффиксами *-ее* и *-ей*, а также формы с приставкой *по-* и без нее для краткости приведены вместе. Заметим, что в среднем формы с суффиксом *-ее* примерно в два раза частотней, чем суффиксом *-ей*, но их соотношение очень отличается у разных прилагательных (в частности, в некоторых случаях форм с суффиксом *-ей* больше). Формы с приставкой *по-* в среднем встречаются в девять раз реже, чем без нее, но их соотношение также сильно варьирует. Так, у прилагательного *тугой* около четверти форм с приставкой *по-*.

Таблица 1. Распределение форм компаратива у прилагательных первой группы без суффикса *-(о)к-*

Прилагательное	**Частотность**	**Кол-во результатов**	**Формы на *-е***	**Формы на *-ее/ей***	**Формы на *-ее/ей* с чередованиями**
богатый	85,0	1873	99,7 %	0,3 %	0,0 %
глухой	40,7	766	99,7 %	0,0 %	0,3 %
густой	47,2	500	98,0 %	1,6 %	0,4 %
крутой	43,4	1809	99,9 %	0,0 %	0,1 %
молодой	414,1	1006	99,5 %	0,5 %	0,0 %
плоский	29,1	686	99,3 %	0,6 %	0,2 %
пологий	5,0	294	98,6 %	0,7 %	0,7 %
простой	275,3	1622	97,1 %	0,2 %	2,7 %
строгий	62,8	693	99,9 %	0,1 %	0,0 %
сухой	83,0	475	95,2 %	0,0 %	4,8 %
твердый	59,5	615	99,0 %	1,0 %	0,0 %
толстый	84,4	1087	99,0 %	0,5 %	0,6 %
тугой	11,9	492	100,0 %	0,0 %	0,0 %
убогий	11,7	342	48,8 %	32,8 %	18,4 %
упругий	10,7	588	68,9 %	27,0 %	4,1 %
частый	114,6	1308	100,0 %	0,0 %	0,0 %
чистый	159,0	1574	99,6 %	0,3 %	0,1 %
яркий	93,9	1701	100,0 %	0,0 %	0,0 %

Таблица 2. Распределение форм компаратива у прилагательных первой группы с суффиксом *-(о)к-*

Прилагательное	Частотность	Кол-во результатов	Формы на *-е*	Формы на *-ее/ей*	Формы на *-ее/ей* с чередованиями	Формы без чередований с выпадением суффикса
близкий	206,8	1505	98,24%	0,0%	0,10%	1,66%
высокий	483,3	1502	99,90%	0,0%	0,10%	0%
глубокий	137,3	1172	99,80%	0,0%	0,20%	0%
низкий	160,7	1197	98,50%	0,0%	0,08%	1,42%
веский	4,9	257	89,50%	0,0%	10,50%	0%
громоздкий	7,1	274	75,91%	14,6%	2,19%	7,30%
дерзкий	8,9	387	60,72%	0%	1,55%	37,73%
жестокий	37,5	581	84,50%	2,4%	13,10%	0%
жуткий	26,1	828	79,38%	1,43%	13, 23%	5,96%
одинокий	44,5	162	48,80%	29,0%	22,20%	0%
скользкий	11,8	325	13,85%	0,92%	3,38%	81,85%
стойкий	10,1	494	92,90%	0,2%	6,90%	0%
тяжкий	26,0	349	97,10%	1,7%	1,20%	0%
хлипкий	2,4	705	96,75%	0,44%	1,48%	1,33%

Анализ данных не выявил статистически значимой зависимости между количеством форм без чередований и такими факторами, как частотность прилагательного или конечный согласный основы, которые оказались значимыми при исследовании глагольных форм. Видно, что проблемы с чередованиями свойственны отдельным прилагательным, которые не объединяет какое-то общее свойство. Возможно, это связано с тем, что интересующая нас группа глаголов на данном этапе развития русского языка активно пополняется, а группа прилагательных — нет. Заметим при этом, что, хотя у многих прилагательных компаративов без чередований в процентном отношении совсем немного, например, 0,5% форм без чередований от такого частотного прилагательного, как *молодой*, — это несколько тысяч результатов.

Также встречаются формы компаратива с выпадением суффикса *-к-*. Данные об этих прилагательных приведены в Таблице 2. Кроме того, мы нашли единичные примеры форм с чередованием, где в противоположность норме выпадает или сохраняется суффикс (*хлиплее, скольже, громо(з)же, жуче, держе, мерже*, а также *близче, низче*). Изредка можно найти и формы с чередованиями, не встречающимися в нормативном русском языке, например, *богаще*, *круще, ярще,* а также крайне редкие примеры типа *скользже*. Наконец, были обнаружены формы *круте, сухе, ярке, близе* и *убоге*. Только первая из них встречается относительно часто (11 однозначных результатов после фильтрации и вычитки), другие крайне редки (1–3 формы на тысячу результатов). Перейдем ко второй группе прилагательных.

Таблица 3. Распределение форм компаратива у сложных прилагательных второй группы и прилагательного *великий*

Прилагательное	Частотность	Кол-во результатов	Формы на *-е*	Формы на *-ее/ей*	Формы на *-ее/ей* с чередованиями
великий	276,3	87	6,9%	20,7%	72,4%
безрукий	1,3	18	0,0%	88,9%	11,1%
близорукий	3,5	26	15,4%	73,1%	11,5%
длинноногий	3,3	114	2,6%	80,7%	16,7%
длиннорукий	0,6	19	0,0%	73,7%	26,3%
жизнестойкий	0,4	16	81,3%	12,5%	6,3%
износостойкий	—	65	70,8%	15,4%	13,9%
легкоплавкий	—	24	83,5%	4,0%	12,5%
лопоухий	1,2	53	1,9%	84,9%	13,2%
морозостойкий	0,6	22	54,6%	18,2%	27,3%
термостойкий	—	21	42,9%	33,3%	23,8%
трудоемкий	4,7	287	74,6%	19,2%	6,3%
тугоплавкий	0,4	43	76,7%	16,3%	7,0%
энергоемкий	1,1	132	89,4%	6,8%	3,8%

Сразу бросается в глаза, что во второй группе значительно больше форм без чередований. Здесь значимое влияние на распределение форм оказал такой фактор: прилагательные *стойкий*, *емкий* и *зоркий* существуют и сами по себе, и от них образуются компаративы с чередованиями, а прилагательных *рукий*, *ногий* и *ухий* нет. Несмотря на то, что варианты соответствующих корней с чередованиями согласных высокочастотны и должны быть на слуху (*ножка, ручка, уши, ушко* и пр.), компаративы с чередованиями от этих сложных прилагательных встречаются редко. В Таблице 4 приведены данные по оставшимся прилагательным второй группы, обозначающим отнесенность к той или иной национальности. Они даются отдельно, так как у многих из них есть формы с выпадением суффикса.

Таблица 4. Распределение форм компаратива у прилагательных, обозначающих отнесенность к той или иной национальности

Прилагательное	Частотность	Кол-во результатов после фильтрации	Формы на *-е*	Формы на *-ее/ей*	Формы на *-ее/ей* с чередованиями	Формы с выпадением суффикса
итальянский	40,9	3	0%	57%	71,43%	14%
испанский	21,1	7	14%	64%	43,00%	0%
немецкий	123,6	27	0%	19%	92,59%	0%
азиатский	10,3	26	4%	57%	11,54%	77%
французский	101,5	244	42,62%	53%	39,34%	15%
русский	530,5	435	4,37%	7%	53,33%	40%

5. Сбор данных в интернете

5.1. Основные принципы и проблемы

Нашей основной задачей было оценить относительную частотность различных форм компаратива, встречающихся в интернете. Мы работали с поисковой машиной «Яндекса». Представим для начала, что мы сравниваем всего две формы, скажем, *суше* и *сухее*. Почему мы не можем просто задать поисковику два запроса и сравнить те числа, которые он выдает на первой странице? Во-первых, потому, что эти числа крайне неточные. Во-вторых, потому, что в результатах будет содержаться много нерелевантных данных, которые необходимо отсеять. Например, *суше* — это не только компаратив, но и форма от существительного *суша*.

Поэтому правильнее задать поисковику запрос, как в примере (2), предполагающий одновременный поиск двух форм, просмотреть какое-то количество полученных результатов и оценить, какова в них относительная частотность этих форм после отсева нерелевантных данных (этот метод изначально был предложен М. А. Холодиловой [2013]).

(2) *"суше" | "сухее"*

Проблема заключается в том, что просмотр результатов — крайне трудоемкое занятие. Их нельзя загрузить в удобном формате и затем подвергнуть первичной автоматической обработке. В связи с этим мы разработали программу «Lingui-Pingui»[3] (https://sites.google.com/site/varyamagomedova) для облегчения задач лингвистического поиска, которая работает через API поисковой машины «Яндекса» (http://api.yandex.ru/) с использованием встроенного языка запросов.

Кроме того, поисковая машина позволяет пользователю просматривать только первую тысячу результатов, однако М. А. Холодилова разработала алгоритм, который позволяет обойти это ограничение [Холодилова 2013: 20-21]. Это делается при помощи ограничения по дате изменения ресурса: для каждой тысячи результатов находится наиболее старый ресурс, и следующий запрос запускается с ограничением по дате так, чтобы результаты каждой следующей тысячи были полностью старше по дате изменения, чем результаты предыдущей. Текст самого запроса при этом остается без изменений. Этот алгоритм воплощен и в нашей программе. Тем не менее, зачастую первой тысячи результатов оказывается вполне достаточно для того, чтобы оценить относительную частотность различных форм.[4]

---

[3] Мы выражаем благодарность Ивану Антонову за неоценимую помощь в разработке и отладке программы, а также Марии Холодиловой за предоставленный алгоритм и помощь в разработке требований к программе.

[4] Как нам подтвердили в службе поддержки «Яндекса», все документы сразу сортируются по релевантности независимо от того, какое слово стоит в запросе первым, а какое последним. Поэтому соотношение найденного соответствует тому, что есть в сети и проиндексировано поисковой системой.

5.2. Программа «Lingui-Pingui»

Программа «Lingui-Pingui» автоматически формирует запросы по заданным параметрам, отсылает их поисковой машине, сохраняет полученные данные и представляет их в удобном для исследователя формате. Кроме того, производится грубая фильтрация данных и считается статистика (общее количество результатов, относительные частотности). Встроенные фильтры и сортировка не заменяют полностью ручной обработки данных, но существенно облегчают эту задачу.

До начала работы программа состоит из скриптов на языке Perl и папки «Input», содержащей файл настроек и текстовые файлы с входными данными. Перед использованием программы необходимо связаться со службой поддержки Yandex.API, чтобы получить возможность отсылать определенное количество поисковых запросов за день (более подробную информацию можно получить на странице http://xml.yandex.ru/ и у службы поддержки). Более детальное описание программы можно найти в работе [Magomedova 2013] или получить, связавшись с нами по электронной почте.

5.2.1. Формирование запросов

«Lingui-Pingui» позволяет автоматизировать создание запросов из заданных списков приставок, корней и суффиксов/окончаний. Рассмотрим работу программы на примере поиска различных форм компаратива от прилагательных *сухой*. Сперва программа формирует по одному запросу на каждый корень, включая туда все возможные комбинации заданных морф. После формирования запроса к нему можно добавить исключения. Скажем, в нашем примере компаратив *суше* совпадает с формами существительного *суша*. Такие данные лучше заранее исключить из поиска. На этом этапе сформированный запрос будет выглядеть следующим образом:

(3) ("сухе" | "сухее" | "сухей" | "посухе" | "посухее" | "посухей" | "суше" | "сушее" | "сушей" | "посуше" | "посушее" | "посушей") ~ *"на суше"* ~ *"по суше"*

Пользователь может задавать запросы и вручную, а также редактировать то, что было создано автоматически. Скажем, в примере (3) комбинация морфов *посухе* не используется в качестве формы компаратива даже в ненормативном языке, и ее можно исключить.

5.2.2. Представление результатов поиска

Сбор данных при поиске ведется по заголовкам страниц, найденных «Яндексом», и по фрагментам текста с этих страниц, которые также выдаются поисковиком. Поэтому в первой тысяче результатов поиска иногда получается больше тысячи строк употреблений искомых форм (если они появляются и в заголовке и во фрагменте текста). Количество слов до и после искомой формы, которые мы берем из исходного фрагмента текста, предоставленного поисковиком, можно задать в файле настроек программы. По умолчанию все фрагменты сохраняются целиком.

Файл со статистикой появляется в подкаталоге «Summary». Там содержится следующая информация:

- количество результатов по версии «Яндекса» («Яндекс» выдает три таких оценки, выбранная нами является наиболее точной и, по крайней мере, в случае, когда количество найденных результатов меньше тысячи и можно посчитать их вручную, совпадает с номером последней найденной страницы);
- количество неповторяющихся фраз;
- количество результатов после отсеивания нестрогих соответствий (поскольку формы, которые мы ищем, — ненормативные, поисковик часто считает их опечатками и включает в результаты поиска соответствующие нормативные формы или другие близкие по написанию слова);
- количество результатов, где одновременно встречаются несколько искомых форм;
- сведения о каждой искомой форме (сколько раз она встретилась и какой это процент от общего числа форм).

5.2.3. Сортировка и последующая обработка результатов

Вкрапления нерелевантных данных в результаты поиска практически неизбежны даже при самом тщательном подходе к формированию запросов. Отфильтровывать такие данные вручную — крайне трудоемкая задача. Автоматическая сортировка существенно ее облегчает.

Для сортировки используются элементы, совместная встречаемость с которыми (в рамках одного фрагмента) практически гарантирует, что найденная форма относится к нужным нам результатам, и элементы, которые почти наверняка являются индикаторами нерелевантных данных. В нашем случае это слова *чем*, *тем*, *гораздо*, *еще* и пр., которые указывают на компаратив, и слова *более*, *менее*, *совсем* и т.д., которые указывают на форму положительной степени с опечаткой, а также, например, буква *і* и слово *“як”*, которые позволяют отсортировать большую часть данных на украинском языке. Программа также содержит инструмент для пересчета статистики для данных, отсмотренных вручную, и инструмент, позволяющий отбирать или исключать из файлов строки, содержащие определенные слова и комбинации, что тоже существенно облегчает ручную обработку данных, так как почти у каждой конкретной формы или конкретного типа опечатки есть свои «маркеры» вместе с которыми именно эта форма(опечатка) часто встречается в результатах.

6. Заключение

Разрушение исторических чередований согласных, исследованное Н. А. Слюсарь и М. А. Холодиловой на материале глаголов, явно затрагивает и формы компаратива. Так как группа прилагательных, от которых образуются нормативные компаративы с чередованиями, не пополняется новыми словами, эти чередования в общем меньше подвержены разрушению, чем у глаголов продуктивного X класса, о которых говорилось в разделе 2. Тем не менее,

у некоторых прилагательных до трети найденных в интернете форм не имеют чередований (это *убогий* и *упругий*, в меньшей степени *громоздкий, лихой, одинокий* и т.д.). Но в полную силу проблемы с чередованиями проявляются, когда люди пытаются образовать синтетический компаратив от прилагательных, у которых нет соответствующих нормативных форм. Мы выбрали для анализа группу таких прилагательных с основами на заднеязычные, так как в русском языке нет ни одной нормативной формы на *-хее*, *-гее* или *-кее*. Тем не менее, встречается множество ненормативных форм такого рода.

В группе сложных прилагательных типа *длинноногий* и *трудоемкий* решающим фактором оказалось то, существует ли вторая часть в виде самостоятельного слова. В этом случае в Интернете находится больше компаративов с чередованием, в то время как от прилагательных на *-рукий*, *-ногий* и *-ухий* они образуются с трудом, несмотря на частотность вариантов соответствующих корней с чередованием (*ручка, ножка, уши, ушко*). Таким образом, на несколько ином материале, чем с глаголами, мы снова можем сделать вывод о том, что всё упирается в доступность конкретной формы от конкретного слова.

Литература

1. *Зализняк А. А.* Грамматический словарь русского языка: Словоизменение. М., 1977.
2. *Князев Ю. П.* Грамматическая семантика: Русский язык в типологической перспективе. М., 2007.
3. *Ляшевская О. Н., Шаров С. А.* Частотный словарь современного русского языка. М., 2009.
4. *Русская грамматика. М., 1980.*
5. *Холодилова М. А.* Позиционные свойства местоимений в русском языке. Магистерская диссертация. СПбГУ, 2013.
6. *Albright A.* Base-driven leveling in Yiddish verb paradigms //Natural Language and Linguistic Theory. Vol. 28. 2010. P. 475–537.
7. *Albright A.* The identification of bases in morphological paradigms. Doctoral dissertation, University of California, 2002.
8. *Anttila R.* Analogy. The Hague, 1977.
9. *Benua L.* Transderivational Identity: Phonological relations between words. Doctoral dissertation, University of Massachusetts, 1997.
10. *Bybee J.* Morphology: A study of the relation between meaning and form. Amsterdam, 1985.
11. *Kiparsky P.* Explanation in phonology. Dordrecht, 1982.
12. *Kiparsky P.* Paradigmatic effects. Stanford,2002.
13. *Kuryłowicz J.* La nature des procèsdits ‘analogiques’ //ActaLinguistica. Vol. 5. 1949. P. 15–37.
14. *Magomedova V.* 2013. Lingui-Pingui a script collection for linguistic search Ms., St.Petersburg State University. (https://sites.google.com/site/varyamagomedova/lingui_yandex/user-guide)

15. *Mańczak W.* Tendences generals des changements analogiques //Lingua. Vol. 7. 1958. P. 298–325, 387–420.
16. *McCarthy J.* Optimal paradigms // Paradigms in phonological theory /ed. by Downing L. J., Hall T. A., Raffelsiefen R. Oxford, 2005. P. 170–210.
17. *Slioussar N., Kholodilova M.* Paradigm leveling in non-standard Russian // Proceedings of the 20th FASL conference. AnnArbot, MI, 2013.

References

1. *Albright A.* Base-driven leveling in Yiddish verb paradigms // Natural Language and Linguistic Theory. Vol. 28. 2010. P. 475–537.
2. *Albright A.* The identification of bases in morphological paradigms. Doctoral dissertation, University of California, 2002.
3. *Anttila R.* Analogy. The Hague, 1977.
4. *Benua L.* Transderivational Identity: Phonological relations between words. Doctoral dissertation, University of Massachusetts, 1997.
5. *Bybee J.* Morphology: A study of the relation between meaning and form. Amsterdam, 1985.
6. *Kholodilova M. A.*, 2013. Pozicionnye svojstva mestoimenij v russkom jazyke. Master Dissertation, St.Petersburg State University.
7. *Kiparsky P.* Explanation in phonology. Dordrecht, 1982.
8. *Kiparsky P.* Paradigmatic effects. Stanford, 2002.
9. *Knjazev, Ju. P.* 2007. Grammaticheskaja semantika: Russkij jazyk v tipologicheskoj perspective (‘Grammatical semantics: Russian language in a typological perspective’). Moscow: Jazyki Slavjanskikh Kul’tur, 2007.
10. *Kuryłowicz J.* La nature des procèsdits ‘analogiques’ // ActaLinguistica. Vol. 5. 1949. P. 15–37.
11. *Ljashevskaja O. H., Sharov S. A.*, 2009. Chastotnyj slovar' sovremennogo russkogo jazyka (na materiale Nacional'nogo korpusa russkogo jazyka). (Modern Russian language frequency dictionary (based on data from Russian National Corpus)). M.: Azbukovnik.
12. *Magomedova V.* 2013. *Lingui-Pingui:* a script collection for linguistic search. Ms., St. Petersburg State University. (https://sites.google.com/site/varyamagomedova/lingui_yandex/user-guide).
13. *Mańczak W.* Tendences generals des changements analogiques // Lingua. Vol. 7. 1958. P. 298–325, 387–420.
14. *McCarthy J.* Optimal paradigms // Paradigms in phonological theory / ed. by Downing L. J., Hall T. A., Raffelsiefen R. Oxford, 2005. P. 170–210.
15. *Shvedova, N.* (ed.). 1982. Russkaja grammatika (‘Russian grammar’). Moscow: Nauka.
16. *Slioussar N., Kholodilova M.* Paradigm leveling in non-standard Russian // Proceedings of the 20th FASL conference. AnnArbot, MI, 2013.
17. *Zaliznyak, A. A.* 1977. Grammaticheskij slovar' russkogo jazyka. Slovoizmenenie (‘The grammatical dictionary of the Russian language. Inflection’). Moscow: Russkij Jazyk, 1977.

МНОГОСТОРОННИЙ ПОДХОД К РЕФЕРЕНЦИИ В АНГЛИЙСКОМ И РУССКОМ ЯЗЫКАХ

МакШейн М. (mcsham2@rpi.edu)

Политехнический институт Ренсселера,
Трой, Нью-Йорк, США

Ключевые слова: семантический анализ, обработка референции, эллипсис

A MULTI-FACETED APPROACH TO REFERENCE RESOLUTION IN ENGLISH AND RUSSIAN

McShane M. (mcsham2@rpi.edu)

Rensselaer Polytechnic Institute, Troy, NY, USA

This paper argues that the detection and resolution of referring expressions can be profitably distributed across modules of a language processing system, rather than being bunched at the end of a text analysis pipeline. The approach is being implemented within the OntoAgent cognitive architecture, which supports the development of multi-functional, language-endowed agents that can collaborate with people in task-oriented applications. Although current development within OntoAgent orients around English, the architecture itself and most of its knowledge bases are language-independent. Drawing upon my past descriptive work on reference and ellipsis in Russian, I will suggest how the same reference resolution strategies might be applied to this and other languages. More generally, I will motivate the need to approach linguistic phenomena in a holistic paradigm, rather than as highly compartmentalized subtasks, which has become the norm for natural language processing applications.

Keywords: semantic analysis, reference resolution, ellipsis

1. Introduction

A common approach in scientific research and technological development is the "divide and conquer" analytical approach, in which the study of naturally-occurring phenomena is divided into subfields and subproblems treated by different research communities, often within different R&D paradigms. Despite well-known scientific and practical motivations for this methodology, there are also drawbacks. For example:

(1) It is often assumed that the prerequisites needed to support the treatment of a (sub)problem are, or will somehow become, available, even if that expectation is unrealistic. In fact, fulfilling the necessary prerequisites is often more difficult than solving the problem itself.

(2) It is assumed that when external prerequisites become available, the solution for the selected subproblem will still be valid. Such speculation is particularly questionable in the case of challenging prerequisites, since the knowledge and processing needed to fulfill them might offer a more natural solution to the original problem.

(3) The solutions for subproblems will ultimately converge into a solution for the full problem. This requires integrating different theories, knowledge bases and input-output expectations of processors, a noble goal fraught with practical, scientific and sociological snares.

(4) It is assumed that some narrowly delineated problem space is a reasonable proxy for the actual problem space, i. e., that methods that are shown to work on a non-real-world problem can be successfully, and without too much additional effort, applied to the solution of the associated real-world problem.

Within the realm of reference resolution, the "divide and conquer" approach has resulted in R&D paradigms that, in my opinion, fail with respect to all of the above points. For example, in Anglo-centered natural language processing (NLP), reference resolution has widely been treated as a machine learning problem specified using the following, oversimplifying rules of the game:

(1) It is expected that manually annotated corpora will be provided for the training and evaluation stages of reference resolution engine development: i. e., reference engines receive as input a perfect syntactic parse of sentences in naturally occurring text, even though achieving this automatically is far beyond the current state of the art.

(2) Referring expressions of interest, called "markables", are manually selected, so the detection stage of reference processing is (artificially) removed.

(3) Markables do not cover all the types of referring expressions—they are limited to cases that lend themselves to resolution using machine learning methods. Specifically:

a. Markables must be noun phrases: referring verbs are excluded.
b. Markables must be overt: ellipsis is excluded.
c. Markables must have overt noun phrase antecedents: verbal, clausal, multi-clause, elided and extra-linguistic antecedents are excluded.
d. The antecedents for markables must be contiguous: dynamically composed sets that can serve as an antecedent for plural referring expressions are excluded.
e. The antecedents for markables must be unambiguous: if annotators are expected to have difficulty agreeing on the antecedent, the given referring expressions is excluded.

Clearly, this problem space—described in far more detail in the Message Understanding Conference (MUC) co-reference task specification (Chinchor 1997)—represents only a fraction of reference resolution challenges present in natural language texts (see McShane 2009 for details).[1] Although work in this paradigm has led to improvements in machine learning methods themselves, I would suggest that this invented problem space has little to do with the actual problems faced by text analysis systems: taking raw text as input and arriving at a full semantic analysis, which necessarily includes the detection and resolution of overt and elided referring expressions. When one addresses the latter problem within an end-to-end language processing system, not only do the challenges look different, so do the available solutions to overcome them.

In this paper, I describe how the detection and resolution of referring expressions is being distributed across modules of the OntoAgent text analysis system. In OntoAgent, individual phenomena are treated as soon as the necessary heuristic evidence becomes available. Our overarching development strategy is to strive for theoretical and practical progress over the long term on the fundamental issues of semantic analysis. As a result, a) we do not expect prerequisites to be fulfilled externally, by systems or human input outside OntoAgent; b) there is no inherent ceiling on the quality of OntoAgent results, and c) we can (and do) exploit OntoAgent in applications even before it reaches its maximum capabilities. Although current applications of OntoAgent orient around English, the architecture itself and most of the supporting knowledge bases are language-independent. Drawing upon my past descriptive work on reference and ellipsis in Russian, I will suggest how the same reference resolution strategies might be applied to this and other languages.

[1] This task, and the manually annotated corpus developed for it, were used to support government-sponsored competitions.

2. Modules of Text Processing

Typically, natural language processing (NLP) systems are implemented as pipelines, such as the one shown in Figure 1.[2]

Fig. 1. A typical NLP pipeline

Although pipeline architectures offer the well-known advantages of modularity, they also suffer from lost potential. For example, key aspects of syntactic analysis, such as decisions about prepositional phrase attachment, cannot be made without semantic input,[3] and key aspects of semantic analysis, such as lexical disambiguation, require reference resolution.[4] In fact, psycholinguistic evidence—which informs but does not constrain our cognitive modeling decisions—shows that people attempt to resolve reference (as shown in eye-tracking experiments) as soon as possible upon hearing elements of input (Tanenhaus 1995). So the "simplifications" promised by a modular architecture can inadvertently lead to the unnatural isolation of phenomena, blocking the necessary, bidirectional passing of heuristic information across modules.

Within the OntoAgent cognitive architecture (described in Section 3), we try to balance the practical utility of a pipeline architecture with the linguistic realities that defy strict compartmentalization. One aspect of this balancing act is distributing the treatment of reference as shown in Figure 2. RR1-RR3 are reference resolution engines that fire before the main reference resolution module. The output of each of these engines can be used to inform all downstream processing.

The cartoon eyeballs in Figure 2 emphasize that the NLP capabilities discussed here are embedded in a more comprehensive agent architecture. This architecture centrally includes static knowledge resources (lexicon, ontology), all manner of processing engines (for simulation, reasoning, NLP), dynamically populated agent memory, and components we will not address here, such as a physiological simulation representing the agent's body. Let us zoom out to this big picture of agent modeling before returning to our topic at hand.

---

[2] Of course, many practical systems do not include all the modules illustrated in the diagram.

[3] For example, in *Shirley hit the boy with the horn*, did Shirley hit a boy holding a horn, or did she use a horn as the instrument with which to hit the boy? This cannot be determined without contextual-semantic analysis.

[4] See McShane et al. 2010 for a discussion of how reference resolution can be used to inform lexical disambiguation.

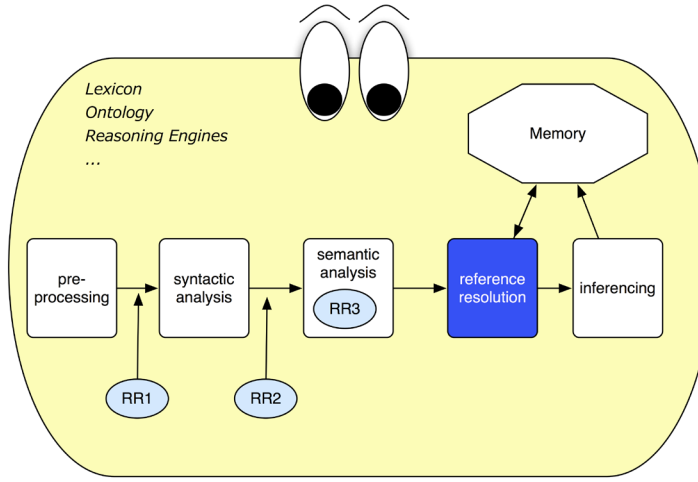


Fig. 2. OntoAgent distribution of reference treatment across processing modules. RR1-RR3 are reference processing engines that are applied before the main "reference resolution" module of text processing.

3. The OntoAgent Environment

OntoAgent is a knowledge-based intelligent agent environment inspired by the traditional goals and motivations of artificial intelligence: attempting to achieve human-level behavior by modeling agents with human-like capabilities of perception, reasoning and action. OntoAgents include integrated physiological and cognitive simulations, with the latter centrally including natural language processing capabilities. A recent prototype application area is Maryland Virtual Patient (Figure 3), a clinician training system in which a cohort of virtual patients can be diagnosed and treated in open-ended, interactive simulations that also optionally include an automatic tutor and additional virtual medical personnel (McShane et al. 2007, 2012 a, b, 2013 a, b; Nirenburg et al. 2008).

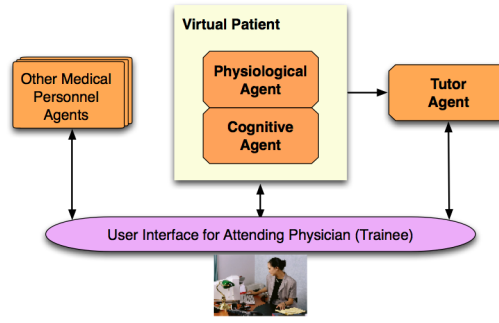


Fig. 3. The Maryland Virtual Patient application

The goal of language understanding in OntoAgent is for the agent to achieve human-level understanding of input text and use that knowledge to populate its memory. Stored memories then support subsequent reasoning and action. In this infrastructure, reference resolution is defined in terms of agent memory: any object or event referred to in a text could be *new* to the agent, in which case a new anchor in memory must be created, or it could be already *known* to the agent, in which case the new information should be appended to the existing anchor. Memory augmentation of this type is presumably what humans achieve, and therefore what intelligent agents must emulate, when processing language input, regardless of whether or not the approaches used in software development show any similarities to the operations of human wetware.

4. Semantic Analysis as the Substrate for Reference

Since reference resolution applies to memory, and since agent memories are stored as ontologically-grounded, disambiguated representations of text meaning, we must begin by briefly describing what we mean by semantic analysis (Nirenburg and Raskin 2004; McShane 2009). Consider example (1), in which [e] indicates an elided category.

(1) Yesterday Sasha played like crazy and [e] fell asleep by eight o'clock.
Вчера Саша играла как зверь и [e] заснула к восьми часам.

This example, like any other, can only have a real-world meaning within some context: i. e., there must be a particular Sasha who, on a particular day, carried out a particular event of playing hard and who was subject to a particular instance of falling asleep at the appropriate interpretation of 8 o'clock, be it a. m. or p. m.

To ground this example and our further discussion of it, let us assume that I am telling this to my mother over the phone. She will know: (i) that Sasha is one of my dogs; (ii) that when she plays she does things like run, fetch her tennis ball, chase squirrels, and wrestle with her canine brother; (iii) which day I'm talking about, by calculating *yesterday* based on the day of the phone call; (iv) that I'm talking about 8 p. m. because only 8 p. m. can occur after a day's worth of playing; and (v) that this is reportable news because Sasha usually doesn't fall asleep by 8:00. If we were to configure a "Mom" intelligent agent with the same background knowledge and reasoning ability as its human counterpart, I would want it to reach these same interpretations.

Two core aspects of arriving at a full semantic interpretation are lexical disambiguation and the establishment of semantic dependencies (Beale and McShane, in preparation). Both of these processes are supported by a text analysis system that relies on a highly detailed computational lexicon and a language-independent ontology. The lexicon includes linked syntactic and semantic expectations for argument-taking words. For example, two of the many senses of the verb *play* are shown below, using a simplified formalism.

```
play-v1
    def                 to play a musical instrument
    example             John is playing (his cello).
    syn-struc
        subject         $var1
        v   $var0
        directobject    $var2 (opt +)
    sem-struc
    PLAY-MUSICAL-INSTRUMENT
            AGENT       ^$var1  (default HUMAN)*
            THEME       ^$var2  (default MUSICAL-INSTRUMENT)*

play-v2
    def             of dogs—to play (fetch balls, run, wrestle, etc.)
    example         Spot is playing in the backyard.
    syn-struc
        subject     $var1
        v           $var0
    sem-struc
        PLAY-DOG
            AGENT       ^$var1  (default DOG)*
```

Play-v1 is optionally transitive and maps to the ontological concept PLAY-MUSICAL-INSTRUMENT. The meaning of the subject (in the basic diathesis) maps to the AGENT case-role, whereas the meaning of the direct object (in a basic diathesis) maps to the THEME case-role.[5] The ontology specifies that the AGENT of PLAY-MUSICAL-INSTRUMENT should be HUMAN, and that the THEME should be MUSICAL-INSTRUMENT. The semantic constraints are marked by an asterisk because they are not actually listed in the lexicon since they are accessible in the ontology. The system will select this sense only if all syntactic and semantic constraints are met. Play-v2, by contrast, is an intransitive sense that maps to PLAY-DOG. It can be selected by the text analyzer only if the subject—to be realized as the AGENT—is of the semantic type DOG.

A core aspect of OntoAgent text processing is that it is largely language independent. In fact, its approach to knowledge-based language processing, which implements the theory of Ontological Semantics (Nirenburg and Raskin 2004), was originally developed for interlingual machine translation. So large portions of a lexicon formulated this way can be ported across languages with only minimal editing required (McShane et al. 2005). For example, Russian equivalents for our examples can be created by simply changing the headwords to играть-v1 and играть-v2. These senses, if processed by a Russian system analogous to our current English one, will generate the same text meaning representations for semantically equivalent input. All subsequent reasoning and action by OntoAgents will then be identical.

---

[5] Non-basic diatheses are treated using syntactic transformations in the analyzer.

The second core static knowledge base is the ontology, which is completely language-independent. Consider a subset of the property values used to describe the concept PLAY-DOG, which describes the special ways that dogs—as opposed to people or horses—play.

```
PLAY-DOG
    AGENT       DOG
    EFFECT      HAPPINESS (RANGE (> .8))
    HAS-EVENT-AS-PART   PLAY-FETCH    (THEME       BALL, STICK, FRISBEE)
                        CHASE         (THEME       SQUIRREL)
                        DIG           (INSTRUMENT  PAW)
                        WRESTLE-DOG
                        RUN
                        SNIFF         (THEME NATURAL-OBJECT)
```

This frame says that dogs are the agents of PLAY-DOG; that the effect of playing is making the dog happy; that typical subevents include fetching a ball, stick or Frisbee, chasing squirrels, digging, wrestling with other dogs, running around, and sniffing the great outdoors. Due to space constraints, many details of ontology form and content omitted; the main point is that the ontology contains key aspects of a person's knowledge about the world, which an intelligent agent should also possess in order to understand language robustly.

The automatic analysis of text, which relies on lexicon and ontology, results in the generation of text meaning representations, which are comprised of numbered instances of ontological concepts connected by ontological relations. The actual numbering of concept instances depends upon the past processing history of a given agent. So, at a given point in its simulated life, an agent (whether processing English or Russian input) might generate the following text meaning representation for our example. Explanatory comments are introduced by semi-colons.

Yesterday Sasha played like crazy and fell asleep by eight o'clock.

```
PLAY-DOG-435
    AGENT           DOG-27
    INTENSITY       1                   ; on the abstract scale {0,1}
    ABSOLUTE-TIME   MONTH 3
                    DAY 13
                    YEAR 2014
    RELATIVE-TIME   < FALL-ASLEEP-271   ; "<" indicates "before"
FALL-ASLEEP-271
    EXPERIENCER     DOG-27
    ABSOLUTE-TIME   HOUR 20
                    MONTH 3
                    DAY 13
                    YEAR 2014
```

RELATIVE-TIME > PLAY-DOG-435

DOG-27

HAS-PERSONAL-NAME Sasha

The concepts themselves and their relationships to other concepts are drawn directly from the sem-struc zones of lexical senses and are combined using the Hunter-Gatherer constraint-based semantic analysis engine (Beale 1996).

This text meaning representation reflects approximately what we expect a human to arrive at when contextually interpreting this input. However, the work of sentence processing is not yet done: the new information has to be incorporated into human or agent memory. Below is a very partial view of what agent memory might look like after hearing this sentence. My human mother—or our "Mom" agent—has an anchor for Sasha that could include hundreds or thousands of pieces of information. The newly reported events in boldface are dynamically added to existing memories. They are also recorded as the heads of their own frames in memory, supplied with all of the extra information available in the text meaning representation above (e. g., absolute and relative times of the events).

DOG-27

HAS-PERSONAL-NAME	SASHA
HAS-FAMILY-NAME	MCSHANE
HAS-OWNER	HUMAN-88
COLOR	BLACK, TAN
WEIGHT	63 (MEASURED-IN POUND)
HAS-BIRTHDATE	(MONTH 8) (DAY 13) (YEAR 2009)
AGENT-OF	INGEST-71, CUDDLE-889, FETCH-204, DIG-336 *[many more events]*, PLAY-DOG-435
EXPERIENCER-OF	STROKE-ANIMAL-44, GROWL-33, *[many more events]*, FALL-ASL**EEP-271**

Population of agent memory is the culmination of contextually grounded semantic analysis, including reference resolution. Earlier, we described, albeit very briefly, how our intelligent agents generate basic text meaning representations, which includes lexical disambiguation and the establishment of semantic dependencies. The next section describes, in slightly more detail, how the processing of reference is distributed across processing modules. Due to constraints of time and space, discussion will focus on those aspects of reference processing that are treated *before* the main reference resolution engine is called.

5. Reference Resolution across Modules

In this section, I describe each of the engines RR1-RR3 (cf. Figure 2) in turn and how they contribute to overall text analysis in our English system. I will also suggest how reference phenomena found in Russian—and, by extension, any other

language—might be treated using the same methods. I should emphasize that I am not orienting around the output of any particular Russian text processors. Instead, I am making certain assumptions about (a) the kinds of inputs and outputs typical of preprocessors and syntactic analyzers cross-linguistically, and (b) the kinds of phenomena that tend to be most challenging for those engines.

5.1. Reference Resolution Engine 1

OntoAgent text processing begins with preprocessing, for which we use the Stanford preprocessor (Klein and Manning 2003a,b). Among its many functionalities are HTML mark-up stripping, sentence and word boundary detection, part of speech tagging, morphological analysis, and certain aspects of named entity recognition. The latter contributes to reference resolution since named entities are referring expressions. However, although the Stanford preprocessor groups named entities, it does not semantically analyze them. It does, however, provide useful input to an OntoAgent engine—a component of RR1—that does semantically analyze them. For example, for the input string *Army Capt. Patrick Horan*, the Stanford preprocessor returns the structure on the left, which OntoAgent uses—along with an onomasticon (lexicon of proper names) and associated rule set—as input when generating the structure on the right:

(NP	HUMAN-1	
(NNP ARMY)	HAS-TITLE	ARMY CAPT.
(NNP CAPT.)	HAS-PERSONAL-NAME	PATRICK
(NNP PATRICK)	HAS-SURNAME	HORAN
(NNP HORAN))		

Why resolve this meaning so early, before "mainline" semantic analysis kicks in? First, because the necessary heuristic evidence is already available; second, because this analysis can then serve as an "island of confidence" for later clause-level semantic analysis. That is, knowing that this individual is HUMAN will help to disambiguate the verb in the clause for which it fills a case role. Named entities in Russian could be treated in the same manner, given a Russian onomasticon and rules about named entity formation comparable to those available for English.

The second aspect of reference processing that requires only the results of preprocessing—along with static knowledge bases that are always accessible—is the detection of certain kinds of ellipsis. For example, when a modal or auxiliary verb occurs before a hard discourse break signaled by a period, colon, semi-colon or question mark[6], this almost always indicates an elided verb, as illustrated by (2):

(2) Zhenya managed to get here on time but Alla couldn't [e].
Женя смог добраться сюда вовремя, а Алла не смогла [e].

[6] Commas are not high-confidence predictors due to widespread cases like the following: Brian didn't, or at least didn't seem to, want to go to Aruba.

Although the inventory of relevant modals/auxiliaries differs across languages, the ellipsis-detection generalization remains the same (McShane et al. 2012d). Moreover, in a highly elliptical language like Russian, it might be useful to seek additional high-frequency, high-confidence elliptical patterns and posit the associated underlying verb. For example, sentences like (3) and (4) could be recognized as elliptical using patterns that refer only to parts of speech (remember, we have not launched syntactic analysis yet).

(3) Noun + preposition + noun + ./? → Noun + [verb] + preposition + noun + ./?
Я в магазин. → Я [e=verb] в магазин.
I to store → I [e=verb] to store
'I'm going to the store.'

(4) Noun + adverb + ./? → Noun + [verb] + adverb + ./?
Ты сразу? → Ты [e=verb] сразу?
You immediately? → You [e=verb] immediately?
'Are you going to do it <come, go, etc.> right now?'

Detecting ellipsis as this stage and positing an underlying verb—even though the system can't know what it means yet—should, presumably, improve syntactic analysis.

5.2. Reference Resolution Engine 2

The next stage of OntoAgent processing is syntactic analysis, for which we use the Stanford syntactic dependency analyzer (de Marneffe et al. 2006). From the parse, OntoAgent detects several types of structures that are potentially elliptical and adds reference-oriented metadata to the current state of analysis to support downstream processing. Two elliptical phenomena that can be detected at this stage are gapping and unexpressed subjects of VP coordinate structures, as illustrated by examples (5) and (6):

(5) Lori had a sandwich and Mary [e], a salad.
Лори съела бутерброд а Мэри [e] салат.

(6) Tom had a sandwich and [e] went to work.
Том съел бутерброд и [e] пошел на работу.

We will concentrate on the example of gapping, though subject ellipsis in these constructions is treated analogously.

The Stanford parser typically analyzes gapping structures as conjoined nominals—essentially, appositives. Simplifying a bit, the output structure for the 2nd half of (5) is: (NP (NP Mary) (NP a salad)). Although this output is incorrect—i.e., the ellipsis is not detected—it is *consistently* and *predictably* incorrect, and this provides us with a useful heuristic. The RR2 engine includes an inventory of clause-level heuristics to evaluate whether a given output with this structure is actually an appositive

or an undetected gapping configuration. If it concludes it is the latter, then the engine: (1) recovers the missing verbal element by copying the verbal string (not yet semantically analyzed!) from the previous conjunct and (2) adds metadata to the copied string that explicitly blocks instance-coreference—i.e., the events in question are of the same type but with different case-role fillers. This revised syntactic output greatly reduces ambiguity in later processing of semantics and reference.

Since Russian widely employs both of these elliptical strategies, this approach to recovering the elided material should be equally applicable—of course, assuming that Russian parsers, like English ones, do not have alternative, high-confidence methods of detecting elided categories.

5.3. Reference Resolution Engine 3

The next stage of processing is basic semantic analysis, which results in text meaning representations like the one described above. The two aspects of reference processing carried out at this stage are (1) the lexically-supported detection of non-referring expressions and (2) the detection and resolution of certain kinds of ellipsis. We will consider each in turn.

Lexically-supported detection of non-referring expressions. Many nouns and verbs are not referring expressions and, therefore, should not be subject to reference resolution procedures. Examples of non-referring expressions in English include, non-exhaustively, pleonastic *it* (*It's cold in here!*); auxiliary uses of polysemous verbs (*I have already finished*); and non-compositional elements of idioms (*He kicked the bucket*, meaning 'He died'). The OntoAgent lexicon includes lexical senses for words and expressions that define the contexts in which they are non-referring and prepare the system to correctly analyze them.

Consider the lexical senses for the English idiom *kick the bucket* and the roughly equivalent Russian *сыграть в ящик*.

```
kick-v4
    def             to die (colloquial, humorous)
    example         Old Mr. Jones kicked the bucket.
    syn-struc
        subject                     $var1
        v                             $var0
        directobject    $var2       (root: bucket (num: sing))
                                    (contains: the (PoS: article))
    sem-struc
        die
            experiencer             ^$var1
    ^$var2      null-sem +

сыграть-v4
    def             умереть (colloquial, humorous)
```

```
example      Старик Иванов сыграл в ящик.
syn-struc
    subject  $var1
    v        $var0
    pp
        prep $var3        (root: в)
        obj  $var2        (root: ящик (num:sing))
sem-struc
die
experiencer  ^$var1
^$var2 null-sem +
^$var3       null-sem +
```

The syn-strucs of both senses indicate which lexical elements must participate in the idiom, along with grammatical constraints on them: in English 'bucket' must be singular and must be used with the article 'the'; in Russian, ящик must be singular and the preposition heading the prepositional phrase must be в. In both languages, the ontological meaning is DIE, and the EXPERIENCER of dying is realized as the subject of the clause. (Stylistic nuances are not central to this discussion and will not be pursued here.) In both languages, the actual words used to express the meaning 'die' are not referring expressions: there is no bucket, box, kicking or playing involved. The sem-strucs indicate that *bucket* and *box,* along with the preposition в in Russian, should not be productively analyzed using the descriptor "null-sem +", which is an abbreivaiton for "null semantics". As such, there will be no explicit trace of these words in the text meaning representations, and reference resolution will not be applied to them, which is exactly what is needed. In short, preparing the system to carry out basic semantic analysis of idioms also blocks the unwarranted search for coreferents for non-referring expressions.

Detection (and resolution) of certain kinds of ellipsis. Just as idioms can be automatically detected and resolved thanks to highly specified lexical entries, so can certain kinds of ellipsis. For example, when modal and aspectual verbs take a direct object that is ontologically an OBJECT, this always indicates semantic ellipsis of the main verb, as in shown in (7) and (8).

(7) Dima desperately wants [e] a dog / a hamburger / a bike.
Дима страшно хочет [e] собаку / гамбургер / велосипед.

(8) Anya finished [e] the article / the blanket only yesterday.
Только вчера Аня кончила [e] статью / плед.

Dedicated lexical senses of modals and aspectuals *anticipate* verbal ellipsis in such contexts and launch calls to procedural semantic routines that attempt resolve the meaning of the unexpressed event. Consider, for example, the lexical sense for *finish* that would be used to analyze (8).

```
finish-v2
    def          to complete some action involving the direct object
```

example Stacy finished the book yesterday
(elided "reading, writing, binding", etc.)
syn-struc
subject $var1
v $var0
directobject $var2
sem-struc
refsem1 (sem EVENT)
AGENT ^$var1
theme ^$var2 (sem OBJECT)
meaning-procedure
seek-specification refsem1 (^$var1 ^$var2)

This sense is used only if $var2 refers to an ontological OBJECT (if it referred to an event there would be no ellipsis, as in *John started washing the car*). When this is the case: (a) there must be an elided EVENT; (b) the meaning of the subject and direct object almost certainly fill the AGENT and THEME case roles, respectively, of that event; and (c) the actual meaning of that event must be dynamically computed based on the meanings of the subject and direct object. The engine that attempts to contextually compute that more specific meaning is triggered by a call to the function "seek-specification", which is listed in the meaning-procedures zone of the entry. If this function can make a confident hypothesis regarding the actual meaning of the event, then that hypothesis is recorded in the text meaning representation—e.g., in the case of Anya finishing a blanket, the hypothesis most strongly suggested by ontological search should be KNIT or CROCHET. If the engine cannot confident constrain the meaning of the event, then the underspecified EVENT remains in the meaning representation. Lexically-recorded procedural semantic routines are used widely in OntoAgent text processing, as described in McShane et al. 2004. The point here is that linguistic expectations about elliptical configurations can be lexically recorded and leveraged to support reference processing *prior* to the working of the dedicated reference module.

Russian employs ellipsis more widely than English, and ellipsis in many configurations can be resolved using highly predictable patterns (McShane 1999, 2000 a, b, 2005). This suggests that anticipatory lexicalization of the patterns could be profitably employed. For example, conjunction structures can involve the ellipsis of the 2nd conjunct's subject and direct object; configurations anticipating such ellipsis can be lexically recorded using the conjunction as a headword.

и-conj12
def: pattern "subject verb direct-object и [e] verb [e]"
ex: Лиза купила пломбир и сразу съела.
Liza bought an ice cream and ate it right up.
comments: a coordinate configuration with ellipsis of the latter subject and direct object

```
syn-struc
    subject         $var1
    v               $var2
    directobject    $var3
    и               $var0
    v               $var4
sem-struc
    CONJUNCTIVE-DISCOURSE-RELATION
    DOMAIN          refsem1
    RANGE           refsem2
meaning-procedure
analyze-clause refsem1 ($var2 (subject:$var1 DO:$var3))
analyze-clause refsem2 ($var4 (subject:$var1 DO:$var3))
```

This lexical sense convers syntactic configurations consisting of a subject, verb, direct object, the conjunction *и,* and another verb, in that order. Within OntoSem, modifiers are always permitted unless explicitly blocked, meaning that a sentence element like сразу 'immediately' in our example will be permitted and compositionally incorporated into the text meaning representation.

The sem-struc says that this configuration contains two clauses that are connected by a conjunctive discourse relation. Those clauses are referred to as refsem1 and refsem1—essentially, pointers to reified structures. Without knowing in advance the particular words that will be used in an input, the lexicon entry cannot predict which concepts will be instantiated or their dependency structure. Those determinations must be made using the normal analysis procedures that are carried out by a pair of calls to the procedural semantic function "analyze-clause".

The key to recovering the elided arguments in this configuration lies in the explicit indication of the arguments of the 2nd verb. Specifically, when the verb indicated by $var2 is being analyzed, it uses $var1 as its subject and $var3 and its direct object; and when the verb indicated by $var4 is analyzed, it *also* uses $var1 as its subject and $var3 as its direct object. In short, this lexical sense anticipates an elliptical structure and explicitly tells the analyzer how to resolve the reference of those elided elements.

Based on my past work on ellipsis in Russian and Polish (cited above), I believe that many elliptical patterns could be effectively treated using this pattern-based strategy. To bypass the rather dense formalism, I will present the patterns via language examples, leaving readers to construe their formal lexical specification independently. Among the relevant patterns are elliptical нет configurations (9), clausal conjunction with an elided 2nd direct object (10), multi-sentence ellipsis of subjects and objects (11), repetition structures, which are often used for emphasis or stylistic effect (12), the ellipsis of verbs of motion and speech (in which the missing verbs may or may not have been detected by RR2) (13), and many more.

(9) Лори любит кататься на велосипеде, а Лиза нет.
Lori likes to ride her bike but Liza doesn't.

(10) «В любом случае завтра, нет, уже сегодня, сменю замок. Здесь этим занимается сторож, куплю **что-нибудь сильно замысловатое**, а он [е] поставит» (Хмелевская).
In any case tomorrow, no, today I'll change the lock. The superintendant deals with those sorts of things. I'll buy something really elaborate and he'll install it.

(11) «Я имел подлость убить сегодня **эту чайку**. Кладу [е] у ваших ног» (Чехов).
«Today I was so base as to kill this seagull. I lay it at your feet» (Chekhov).

(12) «Красное небо, уже начинает восходить луна, и я гнала **лошадь**, гнала [е]» (Чехов).
Red sky, moon on the rise, and I drove that horse on, I drove it hard (Chekhov).

(13) Я не [е] об этом, я [е] о другом.
I'm not talking about that, I'm talking about something else.

To recap, at this point in text processing, the system has generated text meaning representations which already include some reference-oriented decisions: certain elided categories have been detected and a subset of those have been resolved (others await future resolution procedures); in addition, some non-referring expressions have been detected and removed from further consideration by the main reference resolution engine.

Although space does not permit a full description of the interdependencies among processing modules in OntoAgent, one important detail must be mentioned. The lexical senses that support the types of analysis described above are actually leveraged *before* syntactic parsing as well. Specifically, the syntactic patterns recorded in the syn-struc zones of lexicon entries can be used to *force* certain decisions by the syntactic parser. This is particularly important in cases in which the lexically recorded patterns detect ellipsis because, if the parser fails to detect ellipsis, it can produce wildly erroneous output that defies effective semantic analysis. The reason for mentioning this detail out of order with respect to the basic pipeline is pedagogical: the only part of lexical senses important to the parser are syn-struc zones, but it would be strange and unmotivated to describe the syn-struc zones of phrasal lexical senses decoupled from the sem-struc zones of those same entries.

5.4. The Dedicated Reference Resolution Module

Although this paper is centrally about reference resolution, I will not spend much time describing the main reference resolution engine. There are several reasons for this decision: first, constraints of time and space must be observed; second, the associated theory and engine is described in detail in McShane 2012c, 2013c, and Submitted; and third, the real point of this paper is to suggest that reference processing, like any linguistic phenomenon, is best treated holistically rather than in a compartmentalized fashion. However, to avoid a gaping hole in this portrayal of reference processing, I will briefly encapsulate the workings of the dedicated reference module.

This module is called after the basic text meaning representations have been generated. Reference procedures apply to all instances of OBJECTS and EVENTS comprising those meaning representations (recall that all non-referring objects and events will have been excluded by now). The reference engine first determines whether a textual coreferent for an expression should be sought. Although this is straightforward for pronouns and indefinite referring expressions (pronouns always trigger the search for a textual coreferent whereas indefinite referring expressions never do), it is much more complex for verbs, definite referring expressions and named entities. An inventory of knowledge-based algorithms specific to each class of referring expression guides the agent's decision-making with regard to seeking and establishing textual coreference relations. Whether or not an entity has a textual coreferent, it must ultimately be anchored to the agent's memory in the way described earlier, drawing along with it all new information presented in the text. Deciding whether a mentioned entity links to an available anchor in memory or requires the establishment of a new one is carried out based on matching feature value of the input with feature values of available anchors. The success in automating this matching process is highly dependent upon the domain and application. For example, in the Maryland Virtual Patient application, the virtual patient can make certain assumptions about the scope of relevant entities in the world that greatly simplifies the process of memory population management. By contrast, if an OntoAgent is tasked to process a large amount of running text, the challenges of cross-textual reference resolution will surely skyrocket.

The reference resolution module is largely language-independent since its primary source of heuristic evidence is text meaning representations, which are written in the ontological metalanguage. Of course, some aspects of surface realization of language also provide reference clues, such as the distance between a referring expression and each of its candidate antecedents, and, for languages like English, the use of indefinite vs. definite articles. However, much of the heavy-duty reasoning related to memory population and management will be the same for agents operating in any language environment.

6. Concluding Thoughts

The narrow goal of this paper has been to suggest some advantages to distributing reference processing across modules of language processing. Since we are finding this approach useful for English, and since Russian not only shares some difficult reference phenomena with English but adds plenty more to the mix, it seems plausible that this approach might be useful for Russian as well.

However, setting aside the narrow problem of reference resolution, the overarching conclusion is that the big picture of language analysis should more centrally inform both the selection of subproblems by the NLP community and the approaches used to solve them. The "isolationist" mindset that drives much of the recent system building does not show much promise for solving the hard problems of NLP, despite its success in producing engines suitable for simpler tasks supporting limited applications in the near term.

I am not suggesting that the maximally deep, knowledge-heavy semantic analysis pursued by OntoAgents will produce high-quality results over open text in the near term—it certainly will not; after all, its success is predicated on finding solutions to some of the hardest problems in a variety of subareas of cognitive science. Nor am I suggesting that it is inappropriate to build lightweight systems that can solve highly constrained subtasks in the near term; such systems have proved useful for many practical tasks. I am, however, advocating spending more of our collective time thinking, talking and writing about the big picture and building integrated, comprehensive systems because this could fundamentally affect our success in solving individual problems posed by natural language.

References

1. *Beale, S.* 1996. Hunter-Gatherer: Applying Constraint Satisfaction, Branch-and-Bound and Solution Synthesis to Natural Language Semantics. Technical Report, MCCS-96-289, Computing Research Lab, New Mexico State Univ.
2. *Beale, S. & McShane, M.* In preparation. OntoSem language analysis (available upon request).
3. *Chinchor, N.* 1997. MUC-7 Named Entity Recognition Task Definition. Version 3.5, September 17, 1997. Available at http://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html
4. *de Marneffe, M., MacCartney, B., & Manning, C. D.* 2006. Generating typed dependency parses from phrase structure parses. In Proceedings of LREC 2006.
5. *Klein, D., & Manning, C. D.* 2003a. Fast exact inference with a factored model for natural language parsing. Advances in Neural Information Processing Systems 15 (NIPS 2002), Cambridge, MA: MIT Press, pp. 3–10.
6. *Klein, D., & Manning, C. D.* 2003b. Accurate unlexicalized parsing. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics—Volume 1 Pages 423–430.
7. *McShane, M.* 1999. The ellipsis of accusative direct objects in Russian, Polish and Czech. Journal of Slavic Linguistics 7(1): 45–88.
8. *McShane, M.* 2000a. Hierarchies of parallelism in elliptical Polish structures. Journal of Slavic Linguistics, vol. 8, pp. 83–117.
9. *McShane, M.* 2000b. Verbal Ellipsis in Russian, Polish and Czech. Slavic and East European Journal 44(2): 195–233.
10. *McShane, M.* 2005. A Theory of Ellipsis. Oxford University Press.
11. *McShane, M.* 2009. Reference resolution challenges for an intelligent agent: The need for knowledge. IEEE Intelligent Systems, vol. 24, no. 4, pp. 47–58, July/Aug. 2009.
12. *McShane, M.* Submitted. Toward Automating the Resolution of Difficult Referring Expressions.
13. *McShane, M., Beale, S. & Nirenburg, S.* 2004. Some meaning procedures of Ontological Semantics. In: Lino MT, Xavier MF, Ferreira F, Costa R, Silva R, editors. Proceedings of the fourth international conference on language resources and

evaluation (LREC-2004). Conference CD distributed by European Language Resources Association (ELRA), Paris, France.
14. *McShane, M., Nirenburg, S., & Beale, S.* 2005. An NLP lexicon as a largely language independent resource. Machine Translation 19(2): 139–173.
15. *McShane, M., Fantry, G., Beale, S., Nirenburg, S., & Jarrell, B.* 2007. Disease interaction in cognitive simulations for medical training. In: Oxley L., Kulasiri D., editors. MODSIM 2007 International congress on modelling and simulation. Modelling and Simulation Society of Australia and New Zealand, 2007. p. 74–80. Virginia Beach, Sept. 11–13, 2007.
16. *McShane, M., Beale, S., & Nirenburg, S.* 2010. Reference resolution supporting lexical disambiguation. Proceedings of the Fourth IEEE International Conference on Semantic Computing, Carnegie Mellon University, Pittsburgh, PA, Sept. 22–24.
17. *McShane, M., & Nirenburg, S.* 2012a. A knowledge representation language for natural language processing, simulation and reasoning. International Journal of Semantic Computing 6(1).
18. *McShane, M., Beale, S., Nirenburg, S., Jarrell, B., & Fantry, G.* 2012b. Inconsistency as diagnostic tool in a society of intelligent agents. Artificial Intelligence in Medicine (AIIM), 55(3): 137–48.
19. *McShane, M., Nirenburg, S., Beale, S., & Johnson, B.* 2012c. Resolving Elided Scopes of Modality in OntoAgent. Advances in Cognitive Systems, vol. 2, Dec. 2012.
20. *McShane, M., Nirenburg, S., Beale, S. & Johnson, B.* 2012d. Resolving Elided Scopes of Modality in OntoAgent. Advances in Cognitive Systems, vol. 2.
21. *McShane, M., Nirenburg, S., Beale, S., Jarrell, B., Fantry, G., & Mallott, D.* 2013a. Mind-, body- and emotion-reading. Proceedings of IACAP 2013 (International Association for Computing and Philosophy), University of Maryland College Park, July 15–17.
22. *McShane, M., Nirenburg, S., & Jarrell, B.* 2013b. Modeling decision-making biases. Biologically-Inspired Cognitive Architectures (BICA) Journal. Volume 3, January: pages 39–50.
23. *McShane, M., & Nirenburg, S.* 2013c. Use of ontology, lexicon and fact repository for reference resolution in Ontological Semantics. In Oltramari, A., Vossen, P., Qin, L., Hovy, E. (Eds.), New Trends of Research in Ontologies and Lexical Resources. Springer. New Trends of Research in Ontologies and Lexical Resources, Theory and Applications of Natural Language Processing; pp. 157–185.
24. *Nirenburg, S., & Raskin, V.* 2004. Ontological Semantics. Cambridge, Mass.: The MIT Press.
25. *Nirenburg, S., McShane, M., & Beale, S.* 2008. A Simulated Physiological/Cognitive "Double Agent". In: Beal J, Bello P, Cassimatis N, Coen M, Winston P, editors., Papers from the AAAI fall symposium, naturally inspired cognitive architectures, Washington, D. C., Nov. 7–9. AAAI technical report FS-08-06, Menlo Park, CA: AAAI Press.
26. *Tanenhaus M., Spivey-Knowlton M. J., Eberhard K. M., & Sedivy J. C.* 1995. Integration of visual and linguistic information in spoken language comprehension Science; Washington; Jun 16, 1995. Volume: 268 Issue: 5217 Start Page: 1632.

ДУШИ СИРЕНЕВАЯ ЦВЕТЬ… ИЛИ ПРОСТО *КАКАЯ-ТО ХРЕНЬ*? БЕССУФФИКСАЛЬНЫЕ СУЩЕСТВИТЕЛЬНЫЕ В ТЕКСТАХ РУССКИХ ПИСАТЕЛЕЙ[1]

Михеев М. Ю. (m-miheev@rambler.ru)

Научно-исследовательский вычислительный центр
Московского государственного университета
им. М. В. Ломоносова, Москва, Россия

Бессуффиксальные существительные — жен. рода 3-го склонения (*людская молвь*) и — муж. рода 2-го склонения (*конский топ*) еще Пушкиным почитались как вполне законные *коренные* слова русского языка. Друг поэта В. Даль именно за их счет сильно «расширил» свой словарь. На рубеже XIX–XX веков эти слова, особенно первая группа, стали весьма активно употребляться (и порождаться заново!) в текстах русских писателей — как в поэзии, так и в прозе. В частности, их много у Есенина и Шолохова, последний из которых сделал диалектные и просторечные слова отчетливыми маркерами своего стиля.

Ключевые слова: идиостиль, Пушкин, Есенин, Шолохов, Солженицын, маркеры стиля, архаизмы, неологизмы

DUSHI SIRENEVAJA CVET'… OR JUST A NONSENSE (*KAKAJA-TO KHREN'*)? NOUNS WITHOUT SUFFIXS IN THE TEXTS OF RUSSIAN AUTHORS

Mikheev M. Ju. (m-miheev@rambler.ru)

Research Computing Center of Lomonosov Moscow State University, Moscow, Russia

Nouns without suffixes, feminine, the 3rd declination *(e.g. ludskaya molv')* and masculine, the 2nd declination (e.g. *konski top)* were honored by Alexander Pushkin as legitimate *root* Russian words. His friend, Vladimir Dal managed to «expand» his dictionary precisely thanks to these words. At the turn of the XIX–XX century these words, especially the first group, became very frequent in Russian poetry and prose. Some of them were recreated. We can find many interesting examples in Sergei Yesenin's and Mikhail Sholokhov's texts. The latter author, made out of dialect and colloquial words distinct markers of his style.

Keywords: poetic techniques, idiostyle, Pushkin, Yesenine, Sholokhov, Soljenizyn, markers of stile, archaisms, neologisms

---

[1] Статья написана при поддержке гранта РФФИ № 13-06-00402а.

1. Пушкин и — *людская молвь*

Пушкин в Примечании к «Евгению Онегину» (№31 — к строфе XVII главы 5), которое впервые появилось в первом отдельном издании романа, в 1833 году, писал:

В журналах осуждали слова *хлоп*, *молвь* и *топ,* как неудачное нововведение. Слова сии коренные русские. «Вышел Бова из шатра прохладиться и услышал в чистом поле людскую молвь и конский топ» (*Сказка о Бове Королевиче*). *Хлоп* употребляется в просторечии вместо *хлопание*, как *шип* вместо *шипения*:

> *Он шип пустил по-змеиному.* (*Древние русские стихотворения.*)
> *Не должно мешать свободе нашего богатого и прекрасного языка.*[2]

Действительно, слова *шип* и *хлоп* в пору некодифицированности русского языка могли быть понимаемы в значении процесса или действия вместо слов *шипение* и *хлопание*, но — только в просторечии или, во всяком случае, в непринужденном стиле речи. Все же, как мне представляется, в цитированном примечании к роману Пушкин несколько лукавит, выдавая намеренно вызывающую цепь квазинародных словечек, неологизмов и архаизмов (здесь и далее подчеркивания в цитатах везде мои — М. М.) за вполне ординарные слова русского языка того времени:

> *Лай, хохот, пенье, свист и <u>хлоп</u>,*
> *Людская <u>молвь</u> и конский <u>топ</u>!*

С одной стороны, в цитированном примечании 1833 года Пушкин как бы устанавливает знак равенства между нормативным и просторечным вариантами: *шипение — шип*, *хлопание — хлоп*, что на самом деле сейчас не одно и то же (и более того: можно сказать ?*шип сигареты* вместо *шипение*, но невозможно **хлоп дверей* или *— окон*). С другой стороны, Пушкин ссылается на древнюю «Сказку о Бове», т. е. берет словосочетание из нее с *молвью* и *топом* — как цитату, хотя в «Онегине» она без кавычек, но по сути, слова из нее уже в то время скорее были архаизмами, нежели собственно просторечием. В самой «Сказке о Бове» это выражение встречается дважды, в идентичных формулировках, но только порядок компонентов там обратный, ср.:

> (…) *услышал Бова <u>конский топ и людскую молвь</u>* [это в повествовании о Бове, а через несколько страниц текста, уже о его сопернике, потом ставшем слугой или верным помощником, — Полкане:] *услышал <u>конский топ и людскую молвь</u>. И пришел Полкан к Бовину шатру* (…).[3]

---

[2] ФЭБ: http://feb-web.ru/feb/pushkin/texts/push17/vol06/y062001-.htm

[3] «Сказка о Бове Королевиче (Переложение Сергея Сметанина)» — http://dod.ru/surname/bova.htm.

Между тем, надо думать, что упреки «чопорных» критиков, как их тогда, в 1830-е годы окрестил Пушкин, в чем-то тоже оправданны. Поэт пользовался вроде бы вполне законной моделью словообразования для порождения таких «первородных» слов, с нулевым суффиксом, но применял ее, требуя свободы языка несколько «с запасом», огульно. Он настаивал, что полученные в результате варианты равноценны общепринятым. На самом же деле, таких имен существительных, как *хлоп* (в отличие от междометия или глагола *хлоп* или от существительных *хлопóк, хлопание*), *топ* (в отличие от *топот, топание*) и даже *молвь* (в отличие от *молва*), что тогда, так и сейчас нет.[4] Или же — скажем, более аккуратно: их **почти** нет, если исключить сами эти Пушкинские их употребления и вдохновленные им, у других писателей, — из-за чего, конечно, и в словарях теперь они присутствуют.[5] Для русского уха эти слова в соответствующем контексте безусловно понятны, и все-таки почему-то не становятся общеупотребительными, существуя лишь как некие стилистические довески, экзотизмы или поэтизмы: что-то вроде слов на неком близкородственном славянском языке — *славенском* или церковнославянском. Во всяком случае, они ненормативны и пребывают на периферии языка, в зоне цитатных, прецедентных употреблений, на грани авторского казуса или чего-то подслушанного в народной речи.

Примечательно, что еще ранее (предположительно в 1828, но не позднее 1830) Пушкин в своей реакции на критику, обращенную к нему по поводу тех же *молви* и *топа* («Возражение на статью «Атенея»») в запальчивости откликнулся еще более «провокативно», процитировав вначале иронически-издевательское замечание одного из критиков:

> *«Как приятно будет читать роп вм. ропот, топ вм. топот» и проч. На сие замечу моему критику, что роп, топ и проч.*

---

[4] Разве что в современном языке заимствованное *топ* в значении ‘легкая блузка, майка на бретельках’. А вот зато у Есенина будет явный рефлекс Пушкинского (при обращении к буре): *<u>топ</u> громов уйми* (1917).

[5] Число употреблений слова *молвь* в Национальном корпусе русского языка (далее — НК) более 200, объясняется в значительной части нераспознаванием омонимии имени и глагола (роди т. п. = 2 л. ед. ч. *молви*) или тем, что перед нами — явные цитаты *молви* Пушкинской: такова, в частности, Сологубовская «людская молвь» («В толпе» 1907), таковая же и у Салтыкова-Щедрина (1875). Встречаем это словечко и у В. Шишкова, Лескова — с явной стилизацией под народность, в устах соответствующего персонажа: *Патрикей Семеныч, как и со мною у них было, головою понурил, и губа у него одна по другой хлябает, а никакой <u>молви</u> нет*. («Захудалый род» 1874). Есть отдельные примеры у Каткова и Писемского… Но везде слово *молвь* имеет крайне ограниченную сочетаемость. Вот разве что у Гоголя сочетаемость приобретет хоть какую-то свободу: *молвь* у него не только *русская* (хотя и это уже нетривиально, как и в «Дневниках» Цветаевой: *Большевики мне дали хороший русский язык (речь, молвь)…*), но и — *молодецкая*, и *пьяная*! Но это единичные примеры: все-таки за границу цитатного употребления слово, на мой взгляд, так не выходит. Ранее Пушкина, по НК, слово вообще не зафиксировано. Получив от него свой первоначальный толчок, оно займет свое настоящее место — в поэтической речи.

употребляются простолюдимами во многих русских губерниях — NB мне случалось также слышать стукот вместо стук.[6]

Однако даже в словаре Даля, который, вообще-то значительно расширяет реальные границы русского языка своего времени, при наличии таких слов как *стук, стукотня* и *стукотень*, слова **стукот* все же нет, так же как нет и слова **рот* (при наличии *ропота, роптанья* и даже *ропотни*).

2. *Грев и укрыв* — у Даля и Солженицына

Конечно, в русском языке есть большое число вполне законных слов без суффикса, с одной стороны, мужского рода, оканчивающихся на твердый согласный: таких как *навал, обрыв, покров, прибой*… — существительных 2-го склонения, а с другой, — слов женского рода, с основой на мягкий согласный, 3-го склонения, типа *гарь, дочь, нить, постель*. Однако нас интересует группа слов со стилистически экспрессивными дубликатами и то самое пространство «потенциальных» слов, которые претендуют на место в языке, — свободе употребления которых, согласно завету Пушкина, «не до́лжно мешать».

Стилистически маркированный элемент в них, в таких парах, по большей части формально отличается от нейтрального тем, что создается путем простого «укорочения» слова до «первообразного»: если упростить, то путем отбрасывания суффикса (наиболее вызывающе такое сокращение в «хулиганском» примере Пушкина: *ропот* → *рот*: в данном случае часть корня с беглым гласным); хотя, как мы видели и увидим еще, иногда бывает и наоборот, что экспрессивный коррелят возникает как раз путем дополнительного «наращивания» основы: *стук* → *стукот*.[7]

Итак, нас будут интересовать в этой работе именно такие стилистически маркированные слова с нулевым суффиксом, которые в языке имеют суффиксальные стилистически нейтральные дублеты: с одной стороны, такие как *шипение — шип*, *согревание — согрев*,[8] а с другой, такие как *песня — песнь*, *тишина — тишь*, *молва — молвь*…

Это последнее условие только желательное, а не необходимое: чтобы были в наличии и нейтральный, и маркированный стилистически аналог, как для

---

[6] ФЭБ: http://feb-web.ru/feb/pushkin/texts/push10/v07/d07-054.htm. Заметим, что форма слова *простолюдим<u>м</u>* была в то время также одной из возможных — наряду с *простолюдин<u>н</u>*. — Так не сознательно ли здесь Пушкин воспользовался именно ей, как более редкостной, для подчеркивания своей мысли — удержания в языке разнообразия разных вариантов?

[7] Вот вопрос: следует ли считать соотносимыми с обсуждаемой группой слов такие как *горение* (в качестве нейтрального варианта для слова *гарь*) и *дочерь* — как, наоборот, экспрессивного, для *дочь*?

[8] *Согрев* по НК встречается около 30 раз, от Куприна (1915), до наших дней: в основном, в сочетании *для согрева, для согреву* или в еще более экспрессивном варианте *сугрев*, *для сугреву* (частота около 50: начиная от Мельникова-Печерского, 1870, до наших дней).

нормативного *молва* — поэтический архаизм *молвь*, с той же самой приблизительно семантикой (в словаре Даля *молвь* и толкуется попросту через ‘молва’), но с некоторым отличным от нейтрального стилистическим ореолом — в данном случае чего-то архаичного, загадочного, неопределенного…

В начале XX века, сперва как будто только в поэзии, а потом и в прозе активизируются и получают массовое употребление выразительные единицы авторского языка интересующих нас типов. Известно, что Набоков, например, многие свои произведения специально уснащал диковинными словечками, позаимствованными из словаря Даля, чтобы поразить слух и глаз любителя и знатока словесности. Часто они оказывались просто хорошо забытыми перлами из того же Даля. (О языке «вывозного сорта» у Набокова — Михеев 1999:113.)

Позднее — уже в середине XX века — Корней Чуковский будет приветствовать появление таких самобытных слов в прозе Солженицына, критикуя переводы его текста, в которых выразительность народного языка теряется или нивелируется. Имея в виду перевод фразы несобственно-прямой речи повествователя в «Одном дне Ивана Денисовича» (1961), о предстоящей для бригады строителей работе на морозе:

> *Ой, <u>лють</u> там сегодня будет: двадцать семь*
> *с ветерком, ни <u>укрыва</u>, ни <u>грева</u>! —*

Чуковский писал:

> (…) *стиль русского подлинника простонародный, крестьянский: тут и лють, и укрыв, и грев — слова, находящиеся в самом близком родстве со словами молвь и хлоп, к которым в свое время относился с таким сочувствием Пушкин* (Чуковский 1966).

Само слово *лють* (в значении лютость, зверство) — изобретение даже не собственно Солженицынское, а употреблявшееся до него, в частности, А. Ремизовым в 1920-е годы, и еще ранее, в середине XIX в., Ф. Буслаевым. Но к середине XX века оно уже явная экзотика и нужно автору для создания «сказового» повествования, ведущегося как бы от лица крестьянина, деревенского мужика, оказавшегося в лагере. То же касается и слов *укрыв* и *грев* — слов также бессуффиксального типа, но оканчивающихся на твердый согласный. Именно последние, мне кажется, как раз наиболее характерны для Солженицына — в отличие от Шолохова, предпочтение которого было отдано другой модели — бессуффиксальным словам на мягкий согласный.[9] В своем словаре (**СРЯР**) Солженицын указывает *грев* и *грение* в качестве синонимов при общем значении ‘нагревание, согревание, сообщение тепла’, т. е. действие и состояние

[9] Это тем более удивительно, что оба автора происходили из приблизительно одних и тех же диалектных мест: станица Вёшенская, место постоянного обитания Шолохова, — в Ростовской области, а Солженицын жил в Ростове-на-Дону, пока ходил в школу. Возможно тут — полусознательное отталкивание поэтик?

по глаголу *греть* (или, как более систематически описано в словаре Даля, для *грение* и *укрытие* — с пометой 'длительное': так что хочется, сделав следующий вывод, предположить, что *грев* и *укрыв* — это процессы явно более кратковременные). У Даля некоторое существенное различие этих квазисинонимов состоит в том, что *укрывание* — именно 'длит. действие по гл. *укрывать*', т. е. по глаголу несовершенного вида(!), а *укрыв* — действие по гл. *укрыть*, вида совершенного.[10] Но если отвлечься от тонких и почти неуловимых различий в аспектуальных категориях, то слово *укрыв* — это просто экспрессивное словечко, заменяющее слово *укрытие*, такое же, как *грев*, употребленное вместо более нейтральных — *нагревание*, *согревание*, *обогрев* или даже (недостаточно экспрессивного?) — *сугрев* (у Даля — 'печное, избно́е тепло').

Во всем словаре РСЯР мною была выборочно обследована приблизительно десятая часть, а именно слова на букву «О». Среди них оказалось около 200 бессуффиксальных существительных — из которых 110 на твердый согласный (такие как: *обмо́р* — 'продолжит. сост-ие обмершего, похожее на смерть'; *огла́с* — д-е по гл. [*огласить*], *огласка*; *огло́д* — 'объедала; голодный человек'; *о́перед* — д-е по гл. *опередить*). И — уже почти вдвое меньше, 65, существительных на мягкий согласный (*о́бъедь* 'остатки еды, объедки'; *о́держь* 'удерживание, придерживание'; *о́пась* 'опасение, опаска'; *о́ставь* — д-е по гл., оставляемая вещь).

Надо заметить, что практически все эти единицы словаря совпали со словами «живого великорусского языка», которые более чем за 100 лет до этого занес в свой словарь Даль. Хотя оказались и некоторые несоответствия: у Даля, конечно же, таких слов значительно больше (он видимо почти буквально, с неким нордическим педантизмом, подошел к завету своего друга Пушкина по части обеспечения равноправия в языке для таких маргинальных слов, как *молвь, хлоп* и *топ* — наряду с *молвой, хлопанием* и *топотом*, но не *стукотом*), хотя и у Солженицына встречаются изредка такие слова, которых нет в Дале.

3. *Непролазь* М. Шолохова и *усталь* Ф. Абрамова

В работе (Смирнов 2001), посвященной анализу лексики Шолоховского «Тихого Дона», автор фиксировал наличие того, что было им названо — «экспрессивными морфологическими дериватами». Говоря о возможности определить регулярность появления таких элементов в тексте, автор обращает внимание на то, что диалектная лексика и просторечие используются в романе не только для обозначения бытовых реалий — таких как *майдан, баз, курень, зипун* и т. п., но и гораздо более широко, включаясь в реплики персонажей и даже в сами описания от автора. Таким образом весь текст целиком оказывается захвачен диалектной стихией. Смирнов предлагал выделять среди слов, несущих экспрессивную нагрузку в тексте, различные классы: «в процессе первого чтения видно, что они делятся на отдельные вполне устойчивые группы:

---

[10] Таким образом, Даль как будто был склонен усматривать некие аспектуальные различия в формах самих имен существительных?

непролазь, желтень, сухмень, прель, чернь, вызвездь, шелковье, обрубковатый, клешнятый, ногтястый, шишкастый, рукастый, пятниться, прямиться, кучиться, закряжистеть, посмирнеть, заосенеть…» — Здесь первая шестерка слов как раз отвечает нашим требованиям: это существительные 3-го склонения на мягкий согласный, либо с укороченной за счет суффикса основой, либо с удлиненной относительно нормативного слова формой. Автор статьи предлагал в дальнейшем собрать весь словарь Шолоховских «морфологических экспрессем» и, поделив на разновидности, «статистически измерить их распределение по тесту». Однако в указанной статье эта мысль так и не доведена до завершения, а потом продолжения статьи, как было обещано, не последовало.

Ниже приводится такое перечисление — но только первого класса из выделенных, т. е. слов 3-го склонения, либо диалектных, либо просторечных, либо создаваемых автором ad hoc неологизмов, которые рассеяны по всему «Тихому Дону» (далее сокращенно ТД): *белесь* (2+1)[11], *бель* (3), *вызвездь* (1), *гнусь* (2), *голубень* (1), *добычь* (1 — только в черновике), *желтень* (1), *залежь* (ед. ч.) (1), *звень* (1), *калечь* (1), *картофель* — ж. р. в черновике (2), *кипень* (1), *киповень* (1), *марь* (8), *молодь* (2), *наволочь* (1), *немочь* (2), *непроглядь* (3), *непролазь* (4), *ободь* — в значении ‘обод’ (1), *пахоть* (3), *повитель* (10) *повить* (1), *полсть* (21), *по́мочь* (4), *прель* (3), *прожелтень* (2), *прозвездь* (5), *прозелень / сизозелень* (8+1), *промасть* (1), *расторо́пь* (1), *ребятёжь* (1), *ржавь* (2), *ровень* (1), *ростепель* (7), *росшивь* (2), *рыжевень / рыжевель* (2), *сколизь* (3), *стынь* (4), *супонь* (2), *сухмень* (4), *сырь* (1), *темь* (4) *толочь / сутолочь* (1+1), *усталь* (8), *хмарь* (5), *чернь / че́рнеть / исчернь* (8+1+1). Всего же таких экспрессем — более 50, или более 150 всех употреблений, при средней частоте появления (если включить и те, что остались только в черновых вариантах) — на каждой 6-й странице романа.[12] Больше, насколько мне известно, в таком объеме этот прием в литературе никем не эксплуатировался.

Приведу для иллюстрации несколько примеров, из первой части романа:

- реплика героя в диалоге во время рыбалки, в 4 главе:

> *— Обходи глубе! — откуда-то из вязкой черни голос отца. — В 1-м черновом варианте рукописи на этом месте видна правка:* [– Обходи глубе!.. Тина!.. — откуда-то из вязкой ²черной[ты] ¹пустой ~~темноты~~ [голос] ~~ревет~~ отец[ца].][13] *— То есть вместо вязкой черной пустой темноты у автора первоначально появляется вариант вязкой пустой черноты, а затем и — вязкой черни;*

---

[11] В скобках указывается число употреблений данного слова в романе: первоначально *белесь* была равномерно в трех местах: в 1-й, 3-й и 6-й, т. е. последней частях романа, но в позднейших редакциях из 1-й части слово было убрано.

[12] При этом такие слова, как *бестолочь* и *изморось* в этот список, естественно, не включены, как слишком общеупотребительные.

[13] В квадратных скобках — рукописный текст автора с его правкой: зачеркиваниями, исправлениями и вписываниями поверх, по изд. (Рукопись). (Надстрочные цифры — авторские изменения порядка слов.)

- слова от автора, о болезненной свекрови Аксиньи, в 7 главе:

  Посуетившись, падала на кровать и, вытянув в нитку блеклую желтень губ, глядя в потолок звереющими от боли глазами (…). *— В черновике вместо этого еще были:* [блеклые желтые губы]*, но в беловике уже сразу так, как в окончательном тексте, желтень, без видимой правки;*

- при описании погоды, в 8 главе:

  Побрызгивал дождь. Хмарь висела над хутором — в черновике этой фразы еще не было, а в беловике — с правкой: [ˣМарь висела над хутором.] *— где заглавное «Х» явно вписано впереди первоначально заглавного «М»*[14].

В работе (Войлова 1988, 29), посвященной всё той же интересующей нас группе бессуффиксальных существительных 3-го склонения, но выделяемых в прозе Федора Абрамова, отмечалось, что и у этого автора множество таких слов: хотя в целом в современном русском языке нулевая аффиксация не является продуктивным способом словообразования — таким, например, как аффиксальный.[15] Эти слова у писателя носят и литературный характер, и разговорно-просторечный (*темень, хворь, рвань, погань, хмурь, дурь, голь, усталь*) и диалектный — *го́ворь, те́сень, не́роботь, гадь, гребь* (приводятся примеры из текстов «Дом», «Последняя охота», «Пелагея», «Сказание о великом коммунаре»).

В той же работе справедливо отмечен прием расширения смысловой перспективы, или семантического объема слова — как, например, у слова *рвань* за счет совмещения двух значений, прямого и переносного: по Ожегову, 1. разг. ‘рваное платье, обувь’ (…) и 3. перен., собир. ‘мерзавец, негодный человек’. Т.е. сам признак производящей основы здесь как бы усиливается и слово может приобретать значение ‘всё, что связано с рваным’. Или же в слове *хмурь*, где передан и смысл ‘хмурое выражение лица’, и ‘душевный дискомфорт’ (там же, 31-32).

Безусловно, определенную генерализацию способ «нулевой аффиксации» словам придает.

4. *Цветь* и *звень* — Есенина

Видимо, именно конец XIX и начало XX века — время зарождения самой «моды» на употребление таких словечек, которые «сработаны» в народном

---

[14] По Донскому словарю **(ДС)**, *хмара* — туча; *хмарь* — мгла, пелена темных туч; у Даля слова *хмарь* отсутствовало, но было *хмара* — ‘темное облака, туча’, ‘пропасть, бездна’, ‘угрюмый человек’.

[15] Производящими основами, по мнению Войловой, могут быть только основы имен прилагательных и глагола (ср. ниже).

духе. В недавней статье (Плунгян 2013) при обстоятельном рассмотрении истории слова *жуть* в параллель к производящему для него прилагательному *жуткий* было выяснено, что исторически само существительное возникло сравнительно недавно — лишь только в конце XIX века.

Вообще, надо отметить, что интерес к словечкам бессуффиксального типа, оканчивающимся на мягкий согласный, значительно возрос почему-то именно на прошлом рубеже веков. Какому из литературных течений (или кому именно из авторов персонально) русский язык этим обязан, предстоит еще выяснить, но нельзя не обратить внимание на Есенина, который и в поэзии, и в своей ранней прозе (повести «Ярь», 1915), «синтезировал авангардистскую поэтику и образные традиции средневековой русской литературы» (Даньдань 2014: 102), причем контекстом для его неологизмов были «не только лексические эксперименты Хлебникова, Маяковского, Северянина, но и односложные и двусложные безаффиксные существительные, соответствующие архаической фольклорной традиции, как, например, *водь, звень, сочь, трясь* и др.» (там же, с.104).

Ранее об интересующей нас группе слов у Есенина писала и Е. М. Галкина-Федорук (1965, с.20), говоря о словах, образуемых отбрасыванием концов слова от существительного или, чаще, от глагола. При этом она отмечала, что «употребление односложных или двусложных бессуффиксальных существительных и является одной из специфических черт этого общего есенинского тона» (там же, 21), что касалось слов вроде *ширь*, *синь*, *солнь*, *стыть*:

Цветы людей и в <u>солнь</u> и в <u>стыть</u>
Умеют ползать и ходить. (1924)[16]

Исследовательница отмечала у Есенина помимо вполне литературных — таких как *хмарь, крепь, гладь, темь, голь, смоль*, еще и диалектные — *выть* ‘участок пашни, надел земли’ (*Черная, потом пропахшая выть!* 1914), *хлюпь* (1914), *сырь* (1915), *хлябь* и *водь* (1919), *ржавь* (1920), *мреть* (1921)[17], *цветь*

---

[16] Впрочем, слово *солнь* появилось у Есенина с 1916 г., а до него было у Божидара (Богдана Гордеева), в 1914. У Салтыкова-Щедрина ранее уже встречалось *слякоть и <u>стыть</u>* (1857), а у Клюева: *Синь и <u>стыть</u>* (1915).

[17] СРНГ есть только глагол *мреть*, но он очевидно не подходит по смыслу: Перм. ‘смеяться до упада, умирать со смеха’. Вот у Даля этот глагол означает ‘мельтешить, маячить, брезжить, мерцать’, но есть и (он же?) в форме *мре́ять* — т. е. ‘мельтешить, неясно видеться, чуть просвечивать’. Довольно широко употреблялся уже в XIX и начале XX века, у Лескова, Арцибашева, Сергеева-Ценского — Ранее, в XIX в. встречаем употребление этого глагола у Т. Шевченко (*Только чуть <u>мреет</u> влево что-то похожее на лесок*. — «Наймичка» 1844), позднее он же и в поэзии — у Вяч.Иванова (1904). В том числе и в форме причастия — *мреющий* (причастие можно воспринимать как производное от любого из двух глаголов — *мреть* или *мреять*): имя *мреть* у Есенина и образовано от них как бы нерасчлененно: *Все сжигает житейская мреть* (1923). Это можно понять или как ‘утомительное мелькание событий’ (как сказал классик-предшественник, «жизни мышья беготня»), или — ‘что-то застывшее, замершее, умершее’ — если воспринять этот неологизм производным уже от *умирать, обмирать*.

(1924), *никь*,[18] *о́падь* (1925), *стынь* и *звень* (1925) — среди которых много и неологизмов, не всегда, по-видимому, принадлежащих самому Есенину.

Галкина-Федорук утверждала, более того, что «эти слова у нас почти не изучены. Только у Г. Павского отмечен исконно русский характер этих безаффиксных существительных, образованных и от глагола, и от прилагательного» (там же, с.20).[19] И еще очень важное наблюдение: «Эти краткие слова обладают особой ритмичностью из-за корневого ударения, и поэтому бесценны для поэтизации речи».

Важным преимуществом такого существительного без суффиксов — что для Есенина, то и для Пушкина, очевидно, была его выразительность, непривычность и краткость. Приведу замечание Есенина на эту тему, из воспоминаний Горького о нем:

> (…) *Затем, наскоро, заговорил, что глагол «хаять» лучше, чем «порицать». — Короткие слова всегда лучше многосложных, — сказал он.* (М. Горький «О Есенине» 1926.)

Вот что встречается из слов описываемой модели только у одного Есенина: от *трясти — трясь*; от *томить → томь*[20], а также усечением основы прилагательных (*непутевый → не́путь, окольный → око́ль, удобный → удобь, яркий → ярь*.

Но при этом для него вполне возможным оказывается словообразование еще и от имени существительного: от *береза — бе́резь* (*Только видели <u>березь</u> да <u>цветь</u>* — 1924); от *полымя — полымь* (1917); от *воды — водь* (1924), от *ягоды — ягодь* (но это, кажется, в шутку, в инскрипте М. Мурашеву, на книге «Сельский часослов»: *Милому Михаилу на ядреную <u>ягодь</u> слова русского*).

Уже в «Яре» (1915) у Есенина встречаем:

- *дремь* — вместо принятых в языке *дрёма, дремо́та* или диалектного *дремота́* (согласно **НК**, после есенинского «Яра», где это слово встречается дважды, оно также появляется у Е. Замятина и у А. Малышкина, но до него было — в стихах Клюева, 1914);
- *на́вись* — как производное от глагола *нависать / нависнуть* (*черная <u>навись</u> брызнула дождем*) — что очевидно заменяет тривиальное *навес из туч*.

---

[18] Этого слова нет в словарях, так что даже осмыслить без контекста его трудновато. Но по-видимому, его следует возвести к гл. *никнуть*, т. е. нечто поникшему (у Есенина речь идет о ржи, которую топчет конница: (…) *как пошла по ней / Тут рать Деникина, // В сотни верст легла / Прямо в никь она*. («Песнь о великом походе», 1924).

[19] Герасим Петрович Павский (1787-1863), священник, протоиерей, филолог, экзегет, переводчик Библии, основоположник русской библейско-исторической школы, — относил их к «коренным или первообразным именам», которых насчитывал всего около 2 тысяч, а суффиксы он при этом называет «окончательными слогами» (с. 13, 24). (По Павскому всех их следовало бы назвать — «бесчленными именами женского рода»!)

[20] Но в Деулинском (т. е. рязанском) словаре его нет.

В конце 20-х и начале 30-х годов слова с тем же корнем, но еще и с различными приставками активно употреблял С. Кржижановский;[21]

- *лунь*, произведенное от прилагательного *лунный* (или же прямо от — *луна*? Это же слово было ранее у Северянина, так что возможно оно заимствовано Есениным);
- *тужи́ль / тужильный* (по-видимому, тут явное отталкивание от «европейских» слов *траур / траурный*, слишком уж «официально-культурных» для сознания ориентированного на исконность)[22] — производного от глагола *тужить*? (на голове женщины — *тужильная косынка*, в «Яре»)[23].

Широко использует Есенин и диалектные слова — *обро́ть*[24] (1915) и др. Тенденция игры с такого рода оригинальными, «посконными» словечками была и ранее, и остается позднее в стихах Есенина: от *выгибать* — *выгибь* (1918)[25]; от *выбелить* — *выбель* (1921); от *крепить* — *крепь*; от *морозить* — *морозь* (1923), *обморозить* — *о́бморозь* (1915), *поморозить* — *поморозь*); от *омутный* (вместе и *омут*, и нечто *мутное*?) — *о́муть* (1922); от *цвести* — *цветь* (вместо слишком обыкновенного — *цвет* и, по-видимому, слишком обобщенного — *цветение*), от *бредить* — *бредь* (вместо обычного *бред*)[26].

Иногда, впрочем, и у Есенина слова не укорачиваются, а как раз удлиняются: вместо слова с нулевым суффиксом получается более длинное, с наращением суффикса *-ень*: от *голубой* — *голубень*, от *мокрый* вместо обычных *мокрота́* или *мокрая погода* — просторечное *мокре́нь* (а также *мо́креть*, *мокредь* или *мо́кресть*); или от *выбивать* — *выбень*, т. е. ‘нечто выбитое’, как, например, кучи камней на дороге; от *цвет* — *цве́тень* (*В древесную цветень и сочь* — 1925). Или от *желтый* — наряду с литературным *желтизна* — не только просторечно укороченное *желть*, но еще и «удлиненное» *же́лтень*; иногда с добавлением или заменой приставок (последнее слово вместе со словом *про́желтень* — уже Шолоховские, правда, заимствованные им — или просто совпадающие со словечками Андрея Белого) или от *теплый* (*теплеть*), наряду с обычным *оттепель* → еще и *теплынь*, и областное, с наращением уже приставочным — *ростепель*, а также

---

[21] В пяти его текстах за 1928-1933 более 10 раз в формах — *навись, навесь, подвесь, привесь, свесь*.

[22] Фасмер указывает, что это заимствование Петровского времени — из нем. *Trauer* ‘печаль, скорбь, траур’.

[23] Возможно, здесь прилагательное первоначально было образовано от глагола с суффиксом *-л-* по образцу *болелый, ходелый, держалый, озяблый, пахлый* и т. д. (Гецова О. Г. Словообразование // Русская диалектология (под ред. Л. Л. Касаткина). М. 2005, с.107 [Хотя Г. Павский сказал бы, что — от причастия.]

[24] По СРНГ, *о́бороть* — ‘конская узда без удил, недоуздок’ Костр. Арх. Пенз. Ряз. др.; у Ш только как *о́броть* — Моск. Тул. Ряз. Дон. и др. (дважды, но только в ПЦ-1 — в ТД нет).

[25] Позже это станет очень частым словечком С. Кржижановского (9 раз в разных текстах, с 1927 по 1939).

[26] Как описание всего процесса или состояния в целом, а не одного только его внешнего проявления.

с беглым гласным (*ровный → ро́вень*).[27] От прил. *гладкий* — на первом этапе может быть произведено *гладь*, но от последнего еще и *безгла́дь* — *А ныне я в твою безгладь / Пришел, не ведая причины* (как обращение к Кавказу — 1924)[28].

Наконец, возможна тут даже и вставка некоего квазисуффикса: как просторечное производное от *жизнь → жизень* (!) Но что подобная стилизация могла сопровождаться как «укорочением» слова, так и его «удлинением», с наращением, можно видеть на примерах образования слов не только при помощи *-ень*, но и суффиксов *-ость / -есть, -еть,* с близкими «обобщающими» значениями (согласно словарю Т. Ф. Ефремовой): суффикс «*-ен'- / -н'-* (см. также *-вень*) присоединяется к основе мотивирующих инфинитивов, причем гласные финали при этом отсутствуют, а им предшествующая парная твердая согласная чередуется с соответствующей мягкой (как *оползать — оползень, проливаться — проливень*)[29]». Сюда же можно отнести и такие суффиксы, которые отсутствуют «в перечне словообразовательных морфем литературного языка: *-оте́нь, -ове́нь*», *-ома́нь*. Они встречаются в отглагольных существительных, означающих «интенсивное действие» типа *стукотень, хрупотень*[30]*, глухоманъ*.

5. Всякая *хрень* — начала XX века и *смуть* конца XIX

К интересующей нас группе слов закономерно отнести словечко городского просторечия XX века — *хрень* ‘любая вещь, все что угодно, черт знает что, всякая дрянь’, очевидно производное от прилагательного *хреновый*, как бы вмещающее в себя смысл абстрактного существительного *хреновина*, которое, в первую очередь, конечно, просто эвфемизм (*Чего вы, едрена-зелена, уши развесили, всякую хреновину слушаете да еще зубы скалите*? (Артем Веселый. 1924–1932) — оно в ТД 5 раз, но уже было и в ДР, а согласно НК ранее других встретилось у Куприна, еще в 1904). К нему же в параллель, с конца XX в., употребляется и новообразование с удлинением, со словосложением, как бы для пущей выразительности — *хренотень*.

Можно здесь же привести и уже такое выразительное современное и чрезвычайно распространившееся в молодежном жаргоне словечко, как *жесть*,

---

[27] *Ро́вень* — ‘ровность, гладкость’ (СЯШ, ДС, но в СРНГ, со значением ‘ровная поверхность, ровное место, равнина’ — и только как Новг. Пск.). Ну, или исторически еще и чередованием в корне, как в *вопить → выпь* (по Соболевскому и Ильинскому, Фасмер указывает на это слово как связанное чередованием с *вопить, вопль*). В СРНГ из них есть только *дремь* — как ‘дремота’ Тамб. 1912, Сиб. и *не́путь* — ‘легкомысленный, беспутный человек’ Влад. 1853, Нижег. др. (но также и *неп'уть* — наречие в значении сказуемого ‘неладно, нехорошо’ Ряз.).

[28] О неологизмах сходного типа у Есенина и Шолохова — в (Михеев 2014).

[29] Толковый словарь словообразовательных единиц русского языка. М. 1996, с.126–127.

[30] Пожарицкая С. К. Русская диалектология. М. 2005, с.204. Или ср. у Гецовой (о суффиксах *-(о/е)тень, -ень, -овень* и др.): «отвлеченные существительные, мотивируемые глаголами, прилагательными, существительными, для обозначения действия, связанного с шумом или издавания звуков» (указ. соч., с.91).

практически вытеснившее из литературного языка слишком длинные «культурные» слова — *жестокость, жестокосердие, жестоконравие, жестоковыйный* — и потеснившее само слово *жесть* в его исконном значении, как ‘луженое листовое железо’. (Эта *жесть* может использоваться и как выражение одобрения: *Просто жесть!* — в значении *Здорово! Восхитительно! Круто! Вот это да*!)

Или русское ругательство *блядь*, не в своем наиболее распространенном современном значении, а в прежних, по словарю Срезневского — 1) ‘обман’, откуда, по-видимому: 1а) ‘вздор, пустяки, ерунда’, — и в то же время: 2) ‘обманщик’, откуда вероятно: 2а) ‘безумный’ и, собственно, 2б) ‘прелюбодейка’.

Как было сказано, активизация образования подобного рода слов была характерна для начала XX в., особенно его 10-х (в поэзии) и 20-х годов (в прозе): ср. у А. Ремизова *кипь работы* (1917) — вместо более обычных *кипение* или даже диалектного *кипень* ‘кипящая вода, кипяток’ Орл. Ряз.[31]; или у С. Кржижановского — *проступь* (*лазури, улыбки*) (1924, 1925), очевидно, от *проступить*. У последнего автора находим и целую серию слов подобного рода, в разных его текстах, как, например: *вскипь* — от *вскипать*; *налепь* (*афишного столба* — очевидно с ударением на первом слоге); *просыпь* (*снегопада*) (1929) и т. п.

Среди писателей, упражнявшихся в производстве подобных неологизмов, следует упомянуть еще и рано ушедшего (и пожалуй, более известного в свое время, нежели Кржижановский) — Андрея Соболя. Вот из его серии подобных неологизмов: *безлюдь* (вместо — *безлюдье* или *безлюдный, обезлюдеть*), *безмолвь* (от *безмолвие* — видимо, с ударениями в первом случае на приставке, или же просто прибавлением приставки к Пушкинскому *молвь*? а во втором — на основе: *бе́злюдь*, но *безмо́лвь*), а также — *плавь* (от *плавать*), *стужь* (с параллельным *стужа*), *блеснь* (с параллельным *блеснуть*), *затохоль* (или *затухоль*? — *затухать/-нуть*).[32]

Любил словечки такого рода также и Артем Веселый: в его романе «Россия, кровью умытая» (1924–1932) встречаем такой аналог современной *жести*, как *жестель* (по Далю: *жесте́ль* — Ур.-каз. Тамб. ‘ветошь, прошлогодняя трава в поле; жесткая, мерзлая дорога’); а также *неудо́бь* (т. е. земли, неудобные для обработки); *хиль* (по Далю, ‘хворь, болезнь, немочь’ — от *хилый*), *сумь*,[33] *туманъ* (*кисельная туманъ* — при наличии вполне нейтрального в языке слова *туман*) и т. п.

Вполне возможно, что мода на такие слова ведет свое происхождение и с 10-х годов XX в. — от языковых игр «эго-футуристов»: у Игоря Северянина в ходу были вольно-сокращенно-манерные словечки типа *влажь, воль, лунь, хрупь, юнь* (соответственно от *–влажный* вместо *влага/влажность*[34], *воль-*

---

[31] Шолохов добавит сюда еще и *киповень*.

[32] *Пышет печурка второй день, кипит на ней кастрюлька с вином, плавится сахар, корица пряно просится в жаркую плавь; Христианские избы по трубы — по горло — ушли в розовеющую цветень, еврейские домишки более хлопотливые, вылезли тормошливо вперед.* — Все примеры взяты из одного его сборника рассказов «Когда цветет вишня» (1924).

[33] Возможно — производное от *суми́ть*, по Далю, с двумя значениями: ‘образовывать складки, морщины’ и ‘свести с ума’ (так сумасшествие или морщинистость?).

[34] У Кржижановского потом из этого появится: *чавкающая слизь и влажь* (1930).

ный — вместо *воля/вольность, лунный* (ср. ниже), *хрупкий, юный*; у Василия Каменского — даже *бирюзовь, звучаль, зовь, изумрудь* и *раздоль*.[35] Нельзя не отметить искусственности таких словопорождений.

При этом зафиксированное в словарях встречается в языке художественной литературы еще ранее, по крайней мере, такие слова использовались в прямой речи неких «народных» персонажей в прозе, например, у Мамина-Сибиряка: — *От него вся смуть пошла* («Три конца» 1890)[36]. Тут не вполне ясно, что имеется в виду: возможно, *смуть* вместо слова *смута* или же как некая контаминация со словом — *муть*. Но все-таки это уже имеющееся, а не придуманное слово, так как по СРНГ, *смуть*, с пометами Устар. и Обл., то же что *см'утн'я* — т. е. 'раздор, ссора, нелады' Пск. Твер. и др. Однако если писатели «народнического» направления, придерживаясь все-таки реальных фактов, добросовестно заимствовали слова из говоров, то, как представляется, столичные литераторы («золотая молодежь» серебряного века) себя какими-либо рамками не ограничивала.

Всплеск их словотворчества, направленного на слова с нулевым суффиксом, и пришелся как раз на 10–20-е годы XX века. Но сама тяга к словечкам подобного рода существовала и ранее, в XIX-м: способы изображения в беллетристике народной речи традиционно включали в себя как бы намеренное «укорочение» тех слов, которые воспринимались как *интеллигентские*, а потому как бы и инокультурные. Вот у Достоевского: *С глазу на глаз сидим, чего бы, кажется, друг-то друга морочить, комедь играть*? («Братья Карамазовы» 1880)[37]. Ну, или ставшее популярным в просторечии слово *закусь*: *А булки-то у вас чем подмазаны? Нет, уж лучше так, без закуси. Икру-то почем покупаете*? (Н. А. Лейкин. «Его степенство» 1879–1898).

6. *Молчь* и *мертвь* (Пильняк, Кржижановский и др.)

Но вернемся в 20-е годы XX века. Вот у Пильняка очевидно искусственное слово — *мертвь*[38]: от прилагательного *мертвый* или от глагола *мертветь/мертвить* — придуманное, по-видимому, для «оживления» менее экспрессивного

---

[35] *Ты сплошная хрупь, ты вся улыбь*… — эта цитата из Северянина — как бы уже пародийная (К. Чуковский «Эгофутуристы» 1922).

[36] Впрочем далее в глубину я этот вопрос не исследовал: возможно, традиция образования такого рода манерных, «галантерейных» слов существовала и раньше.

[37] Ср. в СРНГ: *коме́дь* — 'о чем-любо очень смешном' Смол. 1858 — вместе с *коме́дья* Твер. 1897 (в ДС нет); *мо́лодь* 'молодой побег растения' Твер. 1855; 'молодая поросль леса' Иркут. 1852, Волог.; 'молодежь' Сл.Ак.Р. 1847 с пометой Стар., Бурят. 1968 (при наличии также *молодьга́* Смол. 1914 и *молод'яжь* Твер. 1936). Или уже современного просторечного *траге́дь* или *про́стынь* — разг. 'простыня' (Большой толковый словарь русского языка. 1-е изд-е: СПб.: Норинт. С. А. Кузнецов. 1998) (в СРНГ нет). Укорочение слов было свойственно и революционной стилистике, что проявилось в словосложении из усеченных основ.

[38] *В каждом человеке, все же, крепко сидит дикарь: эти земли, эта пустыня, эта мертвь — прекрасны, здесь никто не бывал, — так прекрасно и страшно видеть, изведать и знать первый раз*! («Заволочье» 1925).

да и более длинного — *мертвечина* [39]. Возможно, впрочем, неологизм создан и чуть раньше — или независимо –Кржижановским: … *мои первые московские кошмары с их … тупой тоской глухих и мертвых переулков, то подводящих, близко-близко к сиянию и гулу большой и людной площади, то вдруг круто поворачивающих назад в молчь и мертвь, — все эти кошмары, повторяю, в сущности, и были моими первыми сонными ощупями Москвы …* («Штемпель: Москва …» 1925). — Тут же кстати и *молчь*, в которой обращает на себя внимание перекличка — возможно вполне сознательная, этого слова с исходным пунктом нашего описания — Пушкинской *молвью*. Итак, понятно, что имеющая вполне законные традиции построения слова модель делается повышенно востребованной и активно тиражируется (подчас доводясь до самопародии) в первые десятилетия XX века.[40]

С отвагой языкового демиурга перефразировав Пушкинское — *Из мелкой сволочи вербую рать*, Есенин, как известно, дерзко провозгласил следующее:

> *Слова — это граждане, я их полководец, я веду их, мне очень нравятся слова неопределенные, я ставлю их в строй как новобранцев, сегодня они не указаны, а завтра будут в речевом строю такими же, как вся армия»* (из предисловия к сборнику «Стихи скандалиста» 20 марта 1920. Берлин).

И то же самое, по сути, «замутнение» исходного значения — только не такое изящное, как в поэзии, — достигалось в результате языковых операций Шолохова. Солженицын — скорее в стороне от этого пути, избрав для себя, в целях реанимации «живого великорусского» — экзотизмы другого рода, на твердый согласный (очевидно, чтобы не совпасть со своим политическим антагонистом).

Так или иначе, мы видим, что и Есенин, и Шолохов (последний — на десятилетие позднее) практикуют общий прием. Шолохов вполне мог почерпнуть его как из стихов, так и из прозы Есенина, творчество которого было ему близко по духу. Соответствующие маркеры стиля как у первого (в прозаической повести «Яр»), так и у второго, в романе «Тихий Дон», можно рассматривать как серии слов, создающие «народный» колорит, при том что иногда лексика может отходить от исходно родного авторам диалекта (у Есенина — от рязанского, у Шолохова — от донского), включая в себя просторечие и неологизмы, порождаемые обоими в духе литературного сказа — Петра Ершова, Николая Лескова, Алексея Ремизова… Но далее то, что было экспрессемой — диалектным, просторечным выражением или авторским неологизмом, часто пополняет язык — просто как

---

[39] В СРНГ слова *мертвь* нет, в близком к нему значениях есть только *мертво́* и *мертвятина* — ‘мертвечина, падаль’ Новг. 1855; а также *мертвиться* ‘умирать’ Арх.; у Даля еще и *ме́ртвель* — ‘умерший, покойник, побывшийся, умирашка’). Другой пример у того же Пильняка (оттуда же): *нужен был год Арктики, сотни миль дрейфующего льда, умиранье, мертвь, — чтобы скинуть со счетов жизни этот год.*

[40] Она же остается, как будто, актуальной для языка поэзии и в наши дни, в начале XXI-го века: скажем, в стихотворениях Н. Азаровой (об опавших листьях): *перед-лицом / сухая кинь / разноцветная* или: *чопорно прочь / юбилейная млеть*. (Очевидно от *кинуть* и — *молоть*? или же омофон глагола *млеть*?)

его стилистический вариант (как мы видели вначале, в Пушкинских *молви* и *шипе*). Шолохов совершенно очевидно вдохновлялся примером Есенина, что ясно не из заимствований отдельных слов (как раз пословных совпадений у них мало), но главное — из использования самой языковой модели для порождения поэтических архаизмов или диалектизмов.

Итак, вспомним, на чем настаивал Пушкин: «Не должно мешать свободе нашего богатого и прекрасного языка».[41] Пусть в нем умножаются стилистические регистры.

Литература

1. *Войлова К. А.* Семантика и функции имен существительных, образованных безаффиксным способом, в прозе Ф. А. Абрамова // Семантика слова и словоформы в тексте. М.1988.
2. *Галкина-Федорук Е. М.* О стиле поэзии Сергея Есенина. М.: МГУ, 1965.
3. *Гецова О. Г.* Словообразование // Русская диалектология (под ред. Л. Л. Касаткина). М. 2005.
4. *Даньдань У.* (Харбинский педагогический университет). Эстетика авангарда в художественной концепции С. Есенина // Филологические науки 2014.
5. **ДС** — Большой толковый словарь донского казачества. М. 2003.
6. *Ефремова Т. Ф.* Толковый словарь словообразовательных единиц русского языка. М. 1996.
7. *Михеев М.* Заметки о стиле Сирина: еще раз о не-русскости ранней набоковской прозы // Логос № 11/12, М. 1999.
8. *Михеев М.* Переклички «Яра» Есенина с «Тихим Доном» Шолохова — сюжетные, мотивные и текстуальные заимствования // Вестник МГУ. Серия
9. *Филология*. 2014 № 2 [в печати]
10. **НК** — Национальный корпус русского языка — http://ruscorpora.ru/search-main.html
11. *Павский Г. П.* Филологические наблюдения над составом русского языка. [часть] II. Об именах существительных, изд. 2, СПб., 1850.
12. *Плунгян В. А. Жуть* и *жуткий*: от мистицизма к просторечию // Авторская лексикография и история слов: К 50-летию выхода в свет «Словаря языка Пушкина». М.: Азбуковник, 2013.
13. *Пожарицкая С. К.* Русская диалектология. М. 2005.
14. **РСЯР** — Русский словарь языкового расширения. Сост. А. И. Солженицын. М. 2000.

---

[41] Автор статьи благодарен за ее неоднократные обсуждения, с выражением несогласия (весьма ценного), в разное время и в разных формах, следующим лицам: Анне Зализняк, Николаю Перцову и Софье Константиновне Пожарицкой (хотя не на все замечания последней успел среагировать).

15. *Рукопись* — Шолохов М. А. Тихий Дон. Факсимильное издание рукописи. Книга 1 и 2 [Части 1–5: черновые и беловые варианты автографов Шолохова, его жены и свояченицы]. М.-Киев-Париж, 2005.
16. *Смирнов А. А.* Чтения «Тихого Дона». Чтение первое. Предварительный образ стиля (2001, 16 июля) — http://www.philol.msu.ru/~lex/td/?pid=0211&oid=021
17. *Чуковский К.* Живой как жизнь. Высокое искусство. Из англо-американских тетрадей. 1966.

References

1. *Vojlova K. A.* Semantika i funktsii imen sushhestvitel'nykh, obrazovannykh bezaffiksnym sposobom, v proze F. A. Abramova // Semantika slova i slovoformy v tekste. M.1988.
2. *Galkina-Fedoruk E. M.* O stile poehzii Sergeya Esenina. M.: MGU, 1965.
3. *Getsova O. G.* Slovoobrazovanie // Russkaya dialektologiya (pod red. L. L. Kasatkina). M. 2005.
4. *Dan'dan' U.* (KHarbinskij pedagogicheskij universitet). EHstetika avangarda v khudozhestvennoj kontseptsii S. Esenina // Filologicheskie nauki 2014.
5. **ДС** — Bol'shoj tolkovyj slovar' donskogo kazachestva. M. 2003.
6. *Efremova T. F.* Tolkovyj slovar' slovoobrazovatel'nykh edinits russkogo yazyka. M. 1996.
7. *Mikheev M.* Zametki o stile Sirina: eshhe raz o ne-russkosti rannej nabokovskoj prozy // Logos № 11/12, M. 1999.
8. *Mikheev M.* Pereklichki «YAra» Esenina s «Tikhim Donom» SHolokhova — syuzhetnye, motivnye i tekstual'nye zaimstvovaniya // Vestnik MGU. Seriya
9. *Filologiya*. 2014 № 2 [v pechati]
10. **НК** — Natsional'nyj korpus russkogo yazyka — http://ruscorpora.ru/search-main.html
11. *Pavskij G. P.* Filologicheskie nablyudeniya nad sostavom russkogo yazyka. [chast'] II. Ob imenakh sushhestvitel'nykh, izd.2, SPb., 1850.
12. *Plungyan V. A.* ZHut' i zhutkij: ot mistitsizma k prostorechiyu // Avtorskaya leksikografiya i istoriya slov: K 50-letiyu vykhoda v svet «Slovarya yazyka Pushkina». M.: Azbukovnik, 2013.
13. *Pozharitskaya S. K.* Russkaya dialektologiya. M. 2005.
14. **РСЯР** — Russkij slovar' yazykovogo rasshireniya. Sost. A. I. Solzhenitsyn. M. 2000.
15. *Rukopis' — SHolokhov M. A.* Tikhij Don. Faksimil'noe izdanie rukopisi. Kniga 1 i 2 [CHasti 1–5: chernovye i belovye varianty avtografov SHolokhova, ego zheny i svoyachenitsy]. M.-Kiev-Parizh, 2005.
16. *Smirnov A. A.* CHteniya «Tikhogo Dona». CHtenie pervoe. Predvaritel'nyj obraz stilya (2001, 16 iyulya) — http://www.philol.msu.ru/~lex/td/?pid=0211&oid=021
17. *Čukovskij K.* Živoj kak žizn'. Vysokoe iskusstvo. Iz anglo-amerikanskih tetradej. 1966.

К ПРОБЛЕМЕ ОПИСАНИЯ ПРИЛАГАТЕЛЬНЫХ-ИНТЕНСИФИКАТОРОВ ДЛЯ ЗАДАЧ ОБРАБОТКИ ТЕКСТА

Миличевич Я. (jmilicev@dal.ca)
Университет Далхаузи, Галифакс, Канада

Тимошенко С. П. (timoshenko@iitp.ru)
Институт проблем передачи информации
РАН, Москва, Россия

Ключевые слова: лексические функции, интенсификация, модель «Смысл — Текст», толково-комбинаторный словарь, лексическая функция Magn, лингвистический процессор ЭТАП-3, русский, английский

TOWARDS A FINE-GRAINED DESCRIPTION OF INTENSIFYING ADJECTIVES FOR TEXT PROCESSING

Milichevich J. (jmilicev@dal.ca)
Dalhousie University, Halifax, Canada

Timoshenko S. (timoshenko@iitp.ru)
Institute for Information Transmission Problems,
Russian Academy of Sciences, Moscow, Russia

We address collocations of the type "*Intensifying Adjective* + *NOUN*", such as *heavy RAIN* and *complete DISAGREEMENT*, known as Magn type collocations. Such a collocation can be represented as a functional dependency: Magn(*RAIN*) = *heavy*, where Magn is a (lexical) function responsible for the meaning 'very'/'high degree', and *heavy* the value that Magn has with *RAIN*, its keyword. The formalism of lexical functions has proved its usefulness in various NLP tasks, but on close inspection its semantic granularity turns to be insufficient. We propose a refinement of the notion of Magn by distinguishing Magn's semantic subtypes. Our description, which proceeds from the assumption that a choice of a Magn type collocate is not arbitrary, takes into account the following factors:

- semantic class of the keyword (= its semantic label, corresponding to the generic semantic component of its definition) and/or its actants;
- semantic component(s) in the keyword's definition targeted by intensification;
- semantic contrasts observed among Magn type collocates of a given keyword.

We tested our approach on data from the Russian and English explanatory-combinatorial dictionaries developed for the multi-purpose language processing system ETAP-3. As our results show, Magn's semantic subtypes we have identified allow for the encoding of lexicograpahic information in a way that is not only precise but also has predictive power.

Keywords: intensification, lexical function Magn, linguistic processor ETAP-3, Meaning-Text linguistic theory, text processing, Explanatory combinatorial dictionary, collocations, English, Russian

1. The Problem Stated

This paper considers collocations of the type "*Intensifying Adjective + NOUN*", such as *heavy RAIN*, *high PRICE*, *broad DISCUSSION*, *radical REFORM*, etc. More precisely, we investigate the range of meanings that the adjective can carry in such combinations.

Within the Meaning-Text theory, or MTT (Mel'chuk 1974, 2012: 85–159; Kahane 2003), intensifying collocations are described in terms of the lexical function (Wanner, ed., 1996) Magn, a modifier with a very general meaning 'very'/'intense'/'big'; in the dictionary, they are listed (as all collocations) in the entries of their keywords, using the following notation: Magn(*RAIN*) = *heavy*, Magn(*PRICE*) = *high*, and so on.

The fact that the meaning of the lexical function (= LF) Magn is so general and its expression contingent on the keyword has two consequences. On the one hand, semantic contrasts are observed between formally identical elements of its value with different keywords:

(1) ***complete*** *DISAGREEMENT* [≈ 'in **all** aspects of the issue'] vs. ***complete*** *DESPERATION* [≈ 'very **intense**'] vs. ***complete*** *MYSTERY* [≈ 'such that nothing **at all** is known about']

On the other hand, elements of the value of Magn for a single keyword may be less than perfectly synonymous[1]; cf., for instance, the semantically contrasting pairs of boldfaced adjectives in (2a) and (2b), whose simultaneous use in text is not redundant[2]:

(2) a. *Now the House is obviously in* ***complete*** [≈ 'in **all** aspects of the issue'] *and* ***strong*** [strong' ≈ 'such that the views of the parties are **far** apart'] *DISAGREEMENT as to what those safeguards should be.*
b. *This consensus is particularly needed in periods of* ***wide*** [≈ 'having **many** targets'] *and* ***radical*** [≈ 'bringing about a **big** change'] *REFORMS.*

---

[1] Cf. Mel'chuk (1974: 110), translation is ours: "[...] different elements of the value of an LF do not have to be fully synonymous; it is sufficient that they have a common semantic core and that their semantic differences are not "too significant". For the time being, we are unable to give a general answer to the question of the range of allowable semantic differences—i.e., which keyword's collocates that exhibit differences in meaning can be considered as the expressions of the same LF and which ones must be taken to be expressing different LFs.

[2] Our examples of actual LF uses in texts come from Google, Yandex and British National Corpus.

In both cases in (2), meaning differences are due to the fact that the intensification targets different components of the definition of the keyword; more on this will be said later.

A more fine-grained description of intensification is needed for specific language processing tasks, such as machine translation or sense-disambiguation. We are particularly interested in machine translation, where distinguishing "subtypes" of Magn could facilitate lexical choice in cases where there is a mismatch between L_{SOURCE} and L_{TARGET}—i. e., if the intensifiers required/allowed by the keywords that are translational equivalents contrast semantically. We have in mind, for instance, "false friends" like the ones in (3), extracted from the English and Russian dictionaries of the linguistic processor ETAP-3.[3]

(3) Eng. ***hard*** *NEGOTIATIONS* [≈ 'such that the parties are discussing from the positions that are **very different**'] vs. Rus. ***trudnye*** *PEREGOVORY* lit. 'hard negotiations' [≈ 'such that last **long**, the parties sticking to their **very different** respective positions']

The Russian adjective is semantically more complex: it subsumes the meaning of the English one and an additional component, which is expressed in English by a distinct subtype of Magn. This analysis is corroborated by the following examples:

(4) a. $_{Eng}$. *They were* ***hard*** *negotiations, as well as being very* ***long***.
b. $_{Rus}$. $^{??}$*Posle desjati munut* ***trudnych*** *peregovorov, …*
'After ten minutes of hard negotiations, ...'

In English, *hard* and *long* can be simultaneously used with *NEGOTIATION* without any redundancy effect, as is the case in (4a). In Russian, however, using *trudnyj* to qualify negotiations of short duration, as in (4b), is dubious (only four examples have been found on Jandex for 'minutes of hard negotiations', as opposed to several hundreds for 'months of hard negotiations'). The right translation for *hard* in this context is *žostkij* 'tough' (rather than *trudnyj*).

In order to increase the level of granularity of the description and cover the cases such as those illustrated in (2) and (3) above, a technique consisting in indicating the semantic component (within the keyword's definition) targeted by intensification is used. Thus, for the adjectives in (3), we could use the description $\text{Magn}_{['difference']}$, respectively $\text{Magn}_{['difference' \& 'duration']}$. This has been a standard practice within the MTT, applied, for example, in the *Explanatory-Combinatorial Dictionaries* (= ECDs) for Russian (*Mel'chuk, Zholkovsky*, 1984) and French (*Mel'chuk et coll*. 1984–1988–1992–1999) and the *Dicet* lexicographic database for French (*Barque et al*., 2010, *Gader et al*.,

[3] ETAP-3 is a multifunctional NLP environment comprising several applications: a machine translation system, a module of synonymous paraphrasing of sentences, a tagger for syntactic annotation of text corpora, a Universal Networking Language interface, and a computer-assisted language learning tool. Since this environment is largely based on the Meaning-Text Theory, a strongly lexicalist approach to language, it uses as a core component of all the applications a dictionary of a particular type—called *Explanatory-Combinatorial Dictionary*, or ECD; *Apresjan et al*., 2003 offers a brief overview of the ETAP-3 system and the most relevant references.

2012). Here is the gist of the approach, abstracted from (Iordanskaja, Polguère 2005) and (Mel'chuk 2013: 213–215); examples of collocations and their encodings are ours.

Formally speaking, four cases of intensification can be distinguished, based on two independent and combinable parameters:

- The component 'σ' in the definition of the keyword targeted by the intensification: the *central component* (= CC), i. e., the component corresponding to the *genus* in Aristotelian terms, or a *peripheral component* (= PC), i. e., a specific difference component.
- The nature of the link between 'σ' and the intensifying semanteme: *direct* (the intensifying semanteme bears on 'σ') or *indirect* (the intensifying semanteme bears on a component that bears on 'σ'; in Mel'chuk's terms, it is external to the definition of the keyword).

Subtypes 1/2 are "pure" Magn, while subtypes 3/4 additionally contribute "a specific perspective on the situation denoted by the base (i. e., keyword—J. M, S. T.). Typical perspectives are: dimension, duration, quantity, way of doing (emotion, energy, …), etc." (Iordanskaja, Polguère, 2005: 182–183)". The corresponding semantic distinguishers are used as subscripts in the cases where a peripheral component is directly intensified and as superscripts in the cases of indirect intensification.

Table 1. Subtypes of Magn standardly used in the MTT

1) The CC is directly intensified Collocation Keyword's Actantial Structure Meaning of the collocation LF Notation	 radical < sweeping REFORM X's ~ of Y '**big** change$_{CC}$' Magn(REFORM)
2) A PC is directly intensified	 heavy RESPONSIBILITY X's ~ for Y 'X's duty to care for Y_{PC}, Y being **important**' $Magn_{['Y\ is\ important']}$(RESPONSIBILITY)
3) The CC is indirectly intensified	 long <protracted> NEGOTIATIONS X's ~ with Y over Z 'discussion$_{CC}$ whose duration is α, α being **big**' $Magn^{temp}$(NEGOTIATIONS)
4) A PC is indirectly intensified	 wide REFORM X's ~ of Y 'change of Y_{PC} whose number is α, α being **big**' $Magn_2^{quant}$(REFORM)

We think that type 2 Magn too can provide a different perspective on the keyword's meaning; compare, for instance, *heavy responsibility* ('such that Y is important') and *full responsibility* ('such that it rests only on X'), both intensifiers being of type 2. Accordingly, we will consider only type 1 intensifiers to be the "pure" Magn.

While we fully subscribe to the approach outlined above, we believe that it can be enhanced if we take into account another aspect of the keyword meaning, namely its semantic label (Milichevich 1995, Polguère 2003). The semantic label of the lexical unit L corresponds to the central, or generic, component in the definition of L. Semantic labels, such as act, action, communication, event, process, state, artifact, person, etc., are taxonomic characterizers that allow for a compact and formal description of the meaning of lexical units and organize the latter in hierarchical classes, such that each member of a class shares some relevant properties inherited form a higher class. Thus, since processes, states and actions (unlike acts and events) can be characterized for duration and phase, instances of the corresponding semantic labels will by default be characterizable in the same way. Semantic labels are akin to *aspectual classes* of (Vendler, 1967) and Apresjan's *fundamental classes of predicates* proposed in (Apresjan 2006: 75–109). We think that at least some semantic distinguishers used with the LF Magn should be predictable from the semantic label of its keyword and/or from the labels assigned to keyword's semantic actants. To give a concrete example, all instances of the semantic label process (e.g., UNIFICATION, ADJUSTMENT, ADAPTATION, ASSIMILATION) are potentially compatible with two varieties of Magn: $\text{Magn}_{['\text{duration}']}$, with the value *long* <*protracted*>, and $\text{Magn}_{['\text{phase}']}$, with the value *complete*. Conversely, we can predict the (sub-)meaning of the intensifying collocate from the semantic label of its keyword or keyword's actants; for example, *complete* can mean 'at the ultimate stage of' when intensifying $\text{L}_{\text{process}}$, it can mean 'concerning all aspects/elements of' when combining with $\text{L}_{\text{opinion}}$ ([BE IN] AGREEMENT/DISAGREEMENT, APPROVAL/DISAPPROVAL [OF ONE'S ACTIONS], etc.), and so on. While a process has duration and can be conceived as consisting of stages, an opinion is, roughly, information, and is conceived as consisting of elements (cf. *a* ***piece*** *of information*).

Let it be noted that not all labels allow for such inheritance of semantic properties: the higher in the hierarchy a label is, i.e., the more general its meaning, the stronger its predictive power. Nevertheless, the recourse to semantic labels should allow for some interesting generalization, leading to more systematic lexicographic descriptions[4]. Semantic labels were already used in this way in Reuther (1996) and (2003) for a fine-grained description of collocations involving support verbs of Oper_i and Func_i type.

In the next section, we expand on and illustrate our proposal.

---

[4] Another venue to explore when it comes to possible generalizations, mentioned in Iordanskaja & Polguère (2005: 184), are logical links between the structure of the definition and the type of intensification. For example, communications are not gradable, so that they admit only indirect intensification; some standard definition blocks (in definition templates), such as 'potential effect' in the definitions of lexemes denoting some natural phenomena, states, feelings, etc., can also be targeted by intensification (e.g., *devastating STORM, debilitating <life-altering> ILLNESS, petrifying <paralyzing> FRIGHT*, etc.)

2. Our Proposal

We start by briefly evoking the context of the research (2.1) and then proceed to its preliminary findings (2.2).

2.1. Data and Methodology

We used as our primary data source LF-descriptions consigned in the ECD-style English and Russian dictionaries of the ETAP-3 NLP system. These dictionaries make use of four Magn subtypes, encoded as MAGN, MAGN-NS, MAGN-NS1, MAGN-NS2, NS being an abbreviation for "non-standard" (Apresian & Cinman 2002: 117). NS-functions do not have any fast semantic meaning, they are used to fix English-Russian translational equivalents within the range of LF values of the single word. For instance, in the entry for OBSERVANCE, we find: MAGN: COMPLETE1, MAGN-NS: EXACT1, MAGN-NS-1: CLOSE2, because there is a one-to-one correspondence between *complete* OBSERVANCE and *polnoje* SOBL'UDENIJE, *exact* OBSERVANCE and *točnoje* SOBL'UDENIJE, *close* OBSERVANCE and *tščatel'noje* SOBL'UDENIJE. Currently, the number of entries featuring the intensifying LFs in the Russian ECD is as follows: 1841 (MAGN), 461 (MAGN-NS), 99 (MAGN-NS-1), and 42 (MAGN-NS-2). The situation in the English ECD is comparable: 1409 (MAGN), 471 (MAGN-NS), 110 (MAGN-NS-1), and 40 (MAGN-NS-2). The number of actual collocations encoded is much higher, because the intensifying LFs typically yield numerous elements of value for any given keyword. Thus, the CDs provided a wealth of data on which our proposal could be tried out.

As mentioned above, we started from the fact that the semantic label of the lexical unit L is correlated to:

- Possible LFs that L can accept, in our case—subtypes of one specific LF, namely Magn (and, to much lesser extent, the complex LF AntiMagn);
- Possible elements of value for these LFs (or subtypes of a particular LF)[5].

Working with data from the ETAP-3 ECDS, we used as input values of the intensifying LFs (rather than their keywords) and tried to determine whether it is possible to group keywords semantically. For example, consider the adjective ŠIROKIJ 'wide'. As an LF value, it can be translated into English as WIDE, BROAD or LARGE. We extracted from the ETAP-3 dictionaries all the keywords that have these values for Magn, as illustrated in the table below, and tried to find some commonalities in their semantics.

Proceeding from the collocate SHIROKIJ and its translation equivalents, we found groups of semantically similar keywords. For every group we could formulate a refined meaning of Magn expressed by the adjective SHIROKIJ.

---

[5] Cf. Apresjan (2009: 3): "The choice of a particular lexical item L as value of a certain LF from the argument lexeme X is conditioned by a) the nature of the LF in question, b) the lexical meaning of L, and c) the semantic class and subclass of a Vendlerian classification to which X belongs".

Table 2. ŠIROKIJ vs. BROAD and WIDE as values of Magn

Russian lexeme	Magn (Rus.)	Magn (Eng.)	English equivalent	Semantic Class of L and/or L's Semantic Actants (= SemAs)	Target & Type of intensification	Type of Magn
PUBLIKA SPEKTR KLASS	*širokij*	*wide*	AUDIENCE RANGE VARIETY	L = set (of elements)	CC; direct	Magn
OBSUŽDENIJE UČASTIJE DVIŽENIJE PRIZNANIJE	*širokij*	*broad*	DISCUSSION PARTICIPATION MOUVEMENT (*Women's liberation ~ was a struggle for equality.*) RECOGNITION (*They gained ~ for their expertise.)*	L = action & has SemA(s) of type group of people	PC, SemA 1(+2); indirect	$Magn_1^{quant}$
POLNOMOČIJA VOZMOŽNOSTI	*širokij*	*broad*	AUTHORITY POSSIBILITIES	L = property & its 2nd SemA is a plural entity	PC, SemA 2; indirect	$Magn_2^{quant}$
VNEDRENIJE REFORMA	*širokij*	*wide*	IMPLEMENTATION REFORM	L = action & has a SemA of type domain (of activity, etc.).	PC, SemA 2 or 3; indirect	$Magn_2^{quant}$ or $Magn_3^{quant}$
Etc.						

Rus. *širokij* can mean:

With L_{set}:	'containing many elements'	*širokij assortiment* 'wide range'
With L_{action}:	'involving many people'	*širokoje obsuždenije* 'broad discussion'
With $L_{property}$:	'taking into account various elements'	*širokij krugozor* 'broad outlook'
With $L_{physical}$:	'covering a lot of space'	*širokij šag* 'long step'

Of course, good dictionaries usually provide some definitions close to those formulated above. Therefore, our work on the lexical functions and semantic labeling can also be considered as a development of rules for partial disambiguation of adjectives.

An analogous English example—*complete* as a value of Magn can mean:

With $L_{feeling}$:	'intense'	*complete desperation*
With $L_{process}$:	'in the last stage of'	*complete reform*
With $L_{opinion/attitude}$:	'in all relevant aspects'	*complete agreement*
With $L_{property}$:	'the speaker feeling strongly about L'	*complete fool*

Now that we have determined the relevant description factors, the description can also proceed in the opposite direction—starting from the keyword (rather than from values of its intensifier).

Russian examples:

$L_{\text{'interpretation'}}$
IZDEVATEL'STVO$_{\text{X-a nad Y-om}}$ 'mockery', MOŠENNIČESTVO$_{\text{X-a}}$ 'cheating', VREDITEL'STVO$_{\text{X-a}}$ 'sabotage', OSKORBLENIJE$_{\text{X-om Y-a}}$ 'insult', KOMPLIMENT$_{\text{X-a Y-u,}}$ 'compliment', VYGODA$_{\text{X-u ot Ya}}$ 'benefit'

All interpretative[6] lexemes have two major semantic blocks in their definitions: 'X is doing P (with Y)', and 'the Speaker considers P as L'. Thus we can expect some Magn function targeting the degree of speaker's persuasion—how sure he is, to what extent his opinion is strong.

Meaning:	'X is doing P; P is very **typical** L according to the Speaker'
Encoding:	Magn
Expected elements of value:	*prjamoj* 'direct', *nastojaščij* 'true'

$L_{\text{'behavior'}}$
BALOVSTVO 'naughtiness', KAPRIZ 'whim'/'caprice', SHALOST' 'prank', HULIGANSTVO 'hooliganism'

This group is very similar to the previous one, except that the Speaker's evaluation of P is done with respect to some norm. This evaluation is usually embedded in the meaning; e.g., BALOVSTVO denotes a less negative behavior than HULIGANSTVO, and we can expect lexical functions, namely Magn, that target evaluation.

Meaning:	'X is doing P; P is deviating from the norm and the degree of **deviation** is high according to the Speaker'
Encoding:	$\text{Magn}_{\text{evaluation}}$
Expected elements of value:	*bol'šoj* 'big', *ser'oznyj* 'serious'
$\text{AntiMagn}_{\text{evaluation}}$	is also predictable, and this prediction is more accurate.
Expected elements of value:	*melkij* 'minor'

$L_{\text{'relation'}}$
ANALOGIJAY$_{\text{-a (~X)}}$ 'analogy', SOVPADENIEX$_{\text{-a s Y-om}}$ 'coincidence', SHODSTVOX$_{\text{-a s Y-om}}$ 'likeness'

Meaning:	'X and Y are **similar** in many aspects'
Encoding:	$\text{Magn}_{\text{['in all relevant aspects']}}$
Expected elements of value:	*polnyj* 'complete', *točnyj* 'true'

English examples:

L_{set}
AUDIENCE, CROWD, DELEGATION, DEPUTATION, QUEUE

Meaning:	'consisting of (very) many **elements**'
Encoding:	$\text{Magn}_1^{\text{quant}}$
Expected elements of value:	*big*, *large*; *huge*

---

[6] This characterization, as well as the following two ("behavior" and "relation"), are taken form Apresjan's classification of predicates that we already mentioned in Section 1 (Apresjan 2006: 75–109).

In addition, some lexical units have some specific values for this LF; for instance, QUEUE has $Magn_{['length']}$: *long*, targeting the component 'number of people'. (Note the implication relationship between $Magn_1^{quant}$ and $Magn_{['length']}$ of QUEUE; on implicational relations like this, see Conclusion.)

$L_{Y\text{- set}}$
EXHIBITION $_{\sim \text{ by X of Y for Z}}$, SALE $_{\sim \text{ by X of Y to Z for W}}$
Meaning: 'such that **Ys** are many'
Encoding: $Magn_2^{quant}$
Expected elements of value: *big*, *large*; *huge*

$L_{part\ (of)}$
MAJORITY, PERCENTAGE, PORTION
Meaning: 'representing a (very) big **part** of'
Encoding: $Magn_1^{quant}$
Expected elements of value: *big*, *large*
MAJORITY additionally has these elements of value: *vast*, *overwhelming*, while PERCENTAGE has: *huge*.

2.2. Findings

In this subsection, we present two preliminary findings of our study: major subtypes of the LF Magn (2.2.1), and implicative relations existing between these subtypes (2.2.2).

2.2.1. Three Subtypes of Magn

In our analysis of lexicographic data from the ETAP-3 dictionaries, we found three major semantic subtypes of the LF Magn, which could be termed the "pure" Magn, the "aspectual" Magn, and the "emphatic" Magn.

- **"Pure" Magn**

This is the Magn without any additions or shades of meaning, bearing directly on the central semantic component of the keyword's definition (it corresponds to Type 1 Magn in Table 1, Section 1).

- **"Aspectual" Magn**

This Magn subtype provides a specific perspective on the situation denoted by the keyword, in addition to intensifying its meaning (it corresponds to Types 2–4 in Table 1). It does so because it bears on a peripheral component of the keyword's definition and/or by targeting it indirectly, i. e., connecting to the keyword's definition via an intermediate meaning.

- **"Emphatic" Magn**

This Magn subtype has not, as far as we know, been considered before. We became aware of its existence after realizing that some lexemes that combine with the LF Magn and the syntactic negation do not accept the antonymic LF AntiMagn; cf.:

(5) a. *complete/total* [Magn] *ABSURDITY <ANNIHILATION, IGNORANCE, STRANGER>*
b. *not a complete/not a total ABSURDITY <ANNIHILATION, IGNORANCE, STRANGER>*
c. **partial/*slight* [AntiMagn] *ABSURDITY <ANNIHILATION, IGNORANCE, STRANGER>*

It seems that in these cases we are dealing not with the genuine intensification, but rather with rhetorical, or emphatic, one. Two subcases have to be distinguished. The first subcase is represented by nouns like *absurdity, madness, annihilation, destruction* or *bliss*, whose meaning already contains intensification of the highest degree and with which Magn is redundant semantically.[7] What Magn contributes, then, in combination with such a noun, is the Speaker's attitude towards the situation/entity referred to by the noun, something like 'and I feel strongly about this'. The incompatibility of these lexemes with AntiMagn is readily explainable, as well: their high-level internal intensification clashes with the meaning of this LF. The second subcase is represented by nouns such as *ignorance* and *stranger*, with which Magn bears on the internal negation (the component 'absence'/'no') embedded in their definition.[8]

The combinability of a lexeme L with the emphatic Magn should (at least to some extent) be conditioned by L's semantic class. For the time being, we found that "candidate" classes are qualified properties/acts/individuals and destructive acts. More research and more descriptive work is needed in order to better understand the nature of this Magn.

2.2.2. Implicative Relations between Subtypes of Magn

Subtypes of Magn entertain implicative relations with one another, on the one hand, and with the LFs Magn, Bon and Ver, on the other. Here are some examples. If negotiations are hard [$\text{Magn}_{[\text{'difference'}]}$], then we may expect them to be long [$\text{Magn}_{[\text{'duration'}]}$] as well; if an exhibition is large [$\text{Magn}_{2}^{\text{quant}}$], chances are that is also representative [Bon]; if a (socially acceptable) practice is widespread [$\text{Magn}_{1}^{\text{quant}}$], it may well be popular [Bon]; etc. An inversely proportional relationship is possible too: *broad* [$\text{Magn}^{\text{quant}}$] *selection* vs. *fine* [Bon] *selection* (the broader the selection, the less fine it is). While the standard LF notations could be used to describe these implicative relations (something like Qual(*hard*) = *long* | for the keyword negotiations, etc.), it seems that, at least in some cases, interpretation rules based on real-life knowledge would be necessary, as well.

---

[7] Thus, according to LDOCE (*Longman Dictionary of Contemporary English*), *absurdity* is 'the quality of being **completely** stupid or unreasonable', and *annihilation* is 'the fact of destroying something or someone **completely**', *bliss* is '**perfect** happiness or enjoyment', etc.

[8] Cf., again, the way LDOCE explains the meaning of *complete/total/perfect stranger*: used to emphasise that you do not know the person at all.

3. Conclusions

The study of intensifying adjectives allowed us to conclude that, even though we are dealing with restricted lexical co-occurrence, notoriously resistant to generalizations, some degree of generalization can be achieved if we take into account some specific semantic information—namely, semantic label, or class, of intensified lexeme. We did find robust correlations between semantic labels of keywords (and their actants) and Magn subtypes we may expect with them. Not surprisingly, somewhat less reliable is the correlation "L's semantic label ~ elements of value of Magn(L)". Predicting Magn values is easier when they are close to the meaning of the corresponding full adjectives, which is the case of *complete* and *total*, for instance. But in many cases, we simply cannot say with any reasonable certainty what a Magn value could be. Nevertheless, we believe that our proposal constitutes a useful addition to the already existing techniques of description of restricted lexical co-occurrence.

In the future, it would be interesting to establish a fuller set of descriptors to be used with aspectual Magn subtype, working on a larger corpus, and extend the same kind of fine semantic tuning to other LFs.

We will wrap up by mentioning a problem that came up over and over again in the course of our study—namely, the identification of the semantic component in the keyword's definition targeted by intensification. In many cases, we were unable to pinpoint this component. Here is an example. States and processes, being non punctual (and non volitional) facts, can be characterized for duration and phase, and this aspect of their meaning can in principle be targeted by intensification. Now, does the component 'duration' appear at all, at some level of decomposition, in the definitions of lexical units belonging to these classes, or is this a sort of a *semantic quark*, in the sense of Apresjan (1995: 481)? If, as we believe, the second answer is the right one, then the intensification can target some extremely general elements of meaning, such as durativity, distributivity, telicity, volitionality, etc., that do not have an independent lexical expression. As it turns out, these meanings are "distinctive features" in terms of which semantic labels (or classes) are characterized (cf. Milichevich 1995: 69*ff*). Once again, we see the relevance of the concept of semantic label for lexicographic description.

Acknowledgments

We are very grateful to Lidija Iordanskaja and Igor Mel'chuk for their invaluable remarks on a pre-final version of this paper. Many thanks to all the colleagues in the Laboratory of Computational Linguistics at IITP RAS for the warm and creative atmosphere they provided while we worked on this paper. We are especially endebted to Leonid Iomdin, who is "responsible" not only for our acquaintance and the opportunity we got to work together, but also for the inspiration to write this article. Thanks are also due to Tatiana Frolova, whose contribution to lexical functions encoding in the ETAP-3 system can hardly be overestimated.

References

1. *Apresjan Ju.* (1995), An Integrated Description of Language and Systematic Lexicography. Selected Works, vol. 2 [Integral'noye opisanie jazyka i sistemnaja leksikografija. Izbrannye trudy. T. II.], "The Languages of Russian Culture" School [Shkola "Jazyki russkoj kul'tury"], Moscow. [English Translation: Juri Derenick Apresjan, Systematic Lexicography. Oxford University Press, 2000. XVIII p. 304]
2. *Apresjan Ju.* (2006), The Grounds of Systematic Lexicography [Osnovanija sistemnoj leksikografii] in: A Naive Picture of The World and Systematic Lexicography [Jazykovaja kartina mira i sistemnaja leksikografija], Moscow: Jazyki russkoj kul'tury.
3. *Apresjan Ju.* (2009), The Theory of Lexical Functions: An Update, Proceedings of the 4th International Conference on Meaning-Text Theory, Montreal, p. 1–14.
4. *Apresjan Ju., Boguslavskij I., Iomdin L., Lazurskij A., Sannikov V., Sizov V., Tsinman L.* (2003), ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT, in Kahane, S., Nasr, A., eds (2003), Proceedings of the First International Conference on MEaning-Text Theory, Paris, École Normale Supérieure, p. 279–288.
5. *Apresjan Ju., Tsinman L.* (2002), A Formal Model of Sentence Paraphrasing for Text-Oriented NLP systems. [Formal'naja model' perifrazirovanija predlozhenij dlja sistem pererabotki tekstov na estestvennyh jazykah] Russian Language in the scientific coverage [Russkij jazyk v nauchnom osveshhenii] No 4. p. 102–146.
6. *Barque, L., Nasr, A., Polguère, A.* (2010), From the Definitions of the Trésor de la Langue Française To a Semantic Database of the French Language, in Dykstra, A., Schoonheim, T., eds, Proceedings of the XIV Euralex International Congress, Leeuwarden: Fryske Akademy, p. 245–252.
7. *Gader, N., Lux-Pogodlla, V., Polguère, A.* (2012), Hand-Crafting a Lexical Network With a Knowledge-Based Graph Editor, in Proceedings of the Third Workshop on Cognitive Aspects of the Lexicon, CogAlex-III, Mumbai, The COLING 2012 Organizing Committee.
8. *Iordanskaja, L., Polguère, A.* (2005), Hooking up Syntagmatic Lexical Functions to Lexicographic definitions, in Apresjan, Ju., Iomdin, L., eds, Proceedings of the Second International Conference on Meaning-Text Theory, Moscow: Jazyki slavjanskoj kultury, p. 176–186.
9. *Kahane S. (2003), The Meaning-Text Theory, in Angel, V., Eichinger, L. M., Eroms, H.-W., Helwig, P., Heringer, H. J., Lobin, H.,* eds, Dependency and Valency. An International Handbook of Contemporary Research, vol. 1, Berlin/New York: Mouton de Gruyter, p. 456–570.
10. *Mel'chuk I.* (1974), Opyt teorii lingvisticheskih modelej klassa "Smysl—Tekst" [Outline of a Theory of Linguistic Models of the Meaning—Text Type], Moskva: Nauka.
11. *Mel'chuk I.* (2012) Semantics. From Meaning to Text, vol. 1, John Benjamins, Amsterdam/Philadelphia.
12. *Mel'chuk I.* (2013) Semantics. From Meaning to Text, vol. 2, John Benjamins, Amsterdam/Philadelphia.

13. *Mel'chuk I.* et coll. (1984–1988–1992–1999), Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques IV, Montréal: Les Presses de l'Université de Montréal.
14. *Mel'chuk I., Zholkovsky A.* (1984), Explanatory Combinatorial Dictionary of Modern Russian. Semantico-Syntactic Studies of Russian Vocabulary, Vienna: Wiener Slawistischer Almanach, Sonderband 14.
15. *Milićević, J.* (1995), Étiquettes sémantiques dans un dictionnaire formalisé de type "Dictionnaire explicatif et combinatoire" [unpublished master's thesis], Montréal: Université de Montéal.
16. *Polguère, A.* (2003), Étiquetage sémantique des lexies dans la base de données DiCo, t.a.l., 44/2, p. 39–68.
17. *Reuther, T.* (1996), On Dictionary Entries for Support Verbs: The Case of Russian *VESTI*, *PROVODIT´*, *PROIZVODIT´*, in Wanner, ed., 1996: 181–208.
18. *Reuther, T.* (2003), Support Verb Combinations with Existential Verbs (German and Russian), in Kahane, S., Nasr, A., eds (2003), Proceedings of the First International Conference on Meaning-Text Theory, Paris, École Normale Supérieure, p. 1–10.
19. *Vendler, Z.* (1967), Linguistics in Philosophy, Ithaca (NY), Cornell University Press.
20. *Wanner, L.,* ed. (1996), Lexical Functions in Lexicography and Natural Language Processing, John Benjamins, Amsterdam/Philadelphia.

НЕОЛОГИЗМЫ В СОЦИАЛЬНОЙ СЕТИ ФЕЙСБУК

Муравьев Н. А. (nikita.muraviev@gmail.com)

ООО «Лаборатория Цифрового Общества», Москва, Россия; МГУ, Отделение Теоретической и Прикладной Лингвистики, Москва, Россия

Панченко А. И. (a.panchenko@digsolab.com)

ООО «Лаборатория Цифрового Общества», Москва, Россия; Лувенский католический Университет, Лувен, Бельгия

Объедков С. А. (sergei.obj@digsolab.com)

ООО «Лаборатория Цифрового Общества», Москва, Россия; НИУ ВШЭ, Отделение прикладной математики и информатики, Москва, Россия

Исследование проведено на материале наиболее частотных словоформ социальной сети Фейсбук, отсутствующих в словарях. Главная задача исследования состояла в анализе новообразований русского языка с точки зрения наиболее характерных словообразовательных моделей и частей речи, а также адаптации заимствований. Результат нашей работы — словарь из 168 неологизмов и его лингвистический анализ по типу заимствования.

Ключевые слова: неологизм, заимствование, несловарное слово, словообразование, социальная сеть

NEOLOGISMS ON FACEBOOK

Muravyev N. A. (nikita.muraviev@gmail.com)
Digital Society Laboratory LLC, Moscow, Russia;
Moscow State University, Faculty of Theoretical and Applied Linguistics, Moscow, Russia

Panchenko A. I. (a.panchenko@digsolab.com)
Digital Society Laboratory LLC, Moscow, Russia;
Universite catholique de Louvain, Louvain-la-Neuve, Belgium

Obiedkov S. A. (sergei.obj@digsolab.com)
Digital Society Laboratory LLC, Moscow, Russia;
National Research University Higher School of Economics, Department of Applied Mathematics and Information Science, Moscow, Russia

In this paper, we present a study of neologisms and loan words frequently occurring in Facebook user posts. We have collected a dataset of over 573 million posts written during 2006–2013 by Russian-speaking Facebook users. From these, we have built a vocabulary of most frequent lemmatized words missing from the Opencorpora dictionary (http://opencorpora.org/dict.php) the assumption being that many such words have entered common use only recently. This assumption is certainly not true for all the words extracted in this way; for that reason, we manually filtered the automatically obtained list in order to exclude non-Russian or incorrectly lemmatized words, as well as words recorded by other dictionaries or those occurring in pre-2000 texts from the Russian National Corpus (http://www.ruscorpora.ru). The result is a list of 168 words that can potentially be considered neologisms.

We present an attempt at an etymological classification of these neologisms (unsurprisingly, most of them have recently been borrowed from English, but there are also quite a few new words composed of previously borrowed stems) and identify various derivational patterns. We also classify words into several large thematic areas, “internet”, “marketing”, and “multimedia” being among those with the largest number of words.

We consider our results preliminary, but believe that, together with the word base collected in the process, they can serve as a starting point in further studies of neologisms and lexical processes that lead to their acceptance into the mainstream language.

Keywords: neologism, loan word, non-vocabulary, derivation, social network

1. Введение

Систематическое изучение неологизмов и заимствованных слов как отдельной разновидности неологизмов должно дать ответ на различные вопросы, связанные с тем, каким образом меняется лексический состав языка с течением времени и, в частности, по каким моделям и с помощью каких средств интегрируются иноязычные и новообразованные лексические единицы в язык, и как происходит их адаптация. Под моделями мы понимаем как морфологические паттерны словообразования и словоизменения, так его частеречную принадлежность и способность употребляться в различных позициях в предложении.

Помимо формальной стороны вопроса, интерес также представляет семантика лексем, а именно развитие наиболее актуальных семантических полей, лексических отношений внутри них и полисемии лексем. Кроме того, не следует обходить вниманием и такие моменты как, например, вариативность в орфографии заимствованных слов.

Что касается других постоянно пополняющихся классов слов, таких как, например, диалектная лексика, жаргонизмы или имена собственные, они остаются за рамками нашего исследования. Таким образом, объектом нашего исследования являются новообразованные слова, которые необходимо кодифицировать.

Модели словообразования и заимствования представляют не только теоретический, но и прикладной интерес. Так, большинство современных морфологических анализаторов достаточно плохо справляются с обработкой несловарных слов. Знание о том, как образуются новые слова, могло бы помочь в создании более точных систем автоматической обработки текстов. Основной задачей нашей работы было построение первичной словарной базы наиболее частотных неологизмов, используемых в социальных сетях, которая бы стала отправной точкой для подобного рода исследований.

2. История вопроса

Заимствованиям посвящен широкий спектр теоретических работ и типологических работ. В числе фундаментальных трудов по теории заимствованных слов можно назвать классическую работу [Haugen 1950], а также различные исследования последних лет [Capuz 1997, Winter-Froemel 2008, Peperkamp&Dupoux 2003, LaCharité&Paradis 2005]. Существует также немало конкретно-языковых исследований, например, [Hall & Hamann 2003, Volland 1986] по заимствованиям в немецком языке или [Heinemann 2003] — в итальянском языке. Что касается русского языка, то прежде всего стоит упомянуть фундаментальную работу Л. П. Крысина «Иноязычные слова в современном русском языке», а также работы [Брейтер 1997, Дьяков 2003], посвященные англицизмам, и монографию [Маринова 2013]. Неологизмам в узком смысле, то есть словам, образованным от уже имеющихся корней, уделяется несколько меньше внимания. Среди работ по данной тематике можно отметить исследования на материале английского [Fisher 1998], французского [Cougnon 2010]

и каталанского [Estopa 2011] языков. Также можно отметить исследование [Ahmad 1991], посвященное лексическим изменениям в современном английском языке, где рассматриваются как заимствования, так и неологизмы.

3. Анализ несловарных слов социальной сети Фейсбук

3.1. Корпус постов и комментариев русскоязычного Фейсбука

Словарь неологизмов, описанный в данной статье, был построен полуавтоматически из корпуса анонимизированных постов и комментариев Фейсбука[1]. Корпус был получен Лабораторией Цифрового Общества[2] из русскоязычного публичного сегмента социальной сети при помощи программного интерфейса (API) Фейсбука[3]. В набор данных попали сообщения пользователей с «открытым» профилем. Такие профили доступны для чтения для всех других пользователей социальной сети. Построение подобных выборок данных в исследовательских целях распространено в научном сообществе [Catanese et al 2012].

Извлечение неологизмов было произведено из 573 миллионов анонимизированных постов и сообщений 3,2 млн. пользователей социальной сети. В данном эксперименте рассматривались только тексты на русском языке. Язык каждого из входных текстов был определён автоматически при помощи модуля *langid.py* [Lui & Baldwin 2012]. В Таблице 1 приведены основные параметры корпуса текстов русскоязычного Фейсбука, использованного в данной работе. Исследуемый корпус содержит *посты* пользователей за период с 2006 по 2013 год. Первый пост в корпусе датируется 5 августа 2006 года, в то время как последний пост был написан 13 ноября 2013 года. Однако подавляющее большинство постов написано в период с 2011 по 2013 год.

Согласно официальному сайту данной социальной сети [18], всемирная аудитория Фейсбука насчитывала около 1,19 миллиарда активных пользователей в сентябре 2013 года. К сожалению, распределение по странам не представлено на официальной странице. Согласно сайту Internet World Stats [19], количество пользователей Фейсбука в России на конец 2012 года приблизилось к 8 миллионам человек. Таким образом, рассматриваемый набор данных представляет около 40% пользователей русскоязычного Фейсбука 2012 года.

---

[1] http://www.facebook.com

[2] http://www.digsolab.ru

[3] https://developers.facebook.com/tools/explorer

Таблица 1. Статистика корпуса русскоязычных постов и комментариев Фейсбука

Параметр	Значение
Количество анонимизированных пользователей	**3 190 813**
Язык	Русский
Количество постов	426 089 762
Количество комментариев	147 140 265
Количество текстов (посты и комментарии)	**573 230 027**
Количество словоформ в постах	20 775 837 467
Количество словоформ в комментариях	2 759 777 659
Количество словоформ (посты и комментарии)	**23 535 615 126**
Средняя длина поста, словоформ	49
Средняя длина комментария, словоформ	19

3.2. Построение словаря неологизмов

Мы использовали следующий полуавтоматический подход для построения словаря неологизмов из корпуса текстов, описанного выше:

1. Построение частотного словаря. Каждый текст (пост или комментарий) был токенизирован и лемматизирован при помощи морфологического анализатора, основанного на словаре *AOT*[4], который, в свою очередь, основан на словаре Зализняка [Зализняк 1977]. Мы использовали собственную *MapReduce* реализацию модуля построения частотного словаря, основанную на модуле морфологического анализа *RussianMorphology*[5].

2. Морфологическая разметка частотного словаря. Полученный частотный словарь был аннотирован при помощи морфологического анализатора *PyMorphy*[6]. Каждой лемме была присвоена часть речи. Кроме этого, для каждой леммы было указано, входит ли она в словарь *OpenCorpora*[7]. Данный морфологический словарь основан на словаре *AOT* и включает в себя информацию о 388 790 леммах и 5 094 925 словоформах. Построенный словарь несловарных слов Фейсбука был отфильтрован по частоте. К сожалению, на этом этапе произошел отсев новых слов, омонимичных существующим словам русского языка. Примером такого слова, оставшегося за пределами исследования, является используемое в данной статье слово «пост», относительно недавно заимствованное из английского языка в значении «заметка, размещенная пользователем в социальной сети».

---

[4] http://www.aot.ru/

[5] https://code.google.com/p/russianmorphology/

[6] https://bitbucket.org/kmike/pymorphy2

[7] http://opencorpora.org/dict.php

3. Фильтрация словаря экспертами. Полученный частотный словарь несловарных слов оказался крайне «зашумленным». К примеру, в него вошли нерусские слова, неверно лемматизированные слова и другие артефакты автоматической процедуры описанной выше. Следующие несловарные слова оказались среди наиболее частотных в нашем словаре: ть, нибыть, гый, санкт, що, ул, пр, нью, грн, ца, рожение, т.д, від, україни, вебинара, дтпа, кя, свый, плэйкастый, сегода, др, бй, квна, т.е, кг, млма, гр, бо, який, ра, ка, т.к, бть, чи, ск, холти. Для исправления ситуации нами была проведена ручная фильтрация 10 000 наиболее частотных слов. В результате данного этапа был получен словарь, из которого мы отобрали 624 наиболее частотных несловарных слова[8].

4. Лингвистическая фильтрация словаря. В результате фильтрации на предыдущем этапе был получен гораздо менее «зашумленный» список несловарных слов, которые часто употребляются пользователями социальной сети. Однако в данный список попало большое количество несловарных слов, которые нельзя отнести к неологизмам. Во-первых, в список попало много имен собственных: географические объекты, имена, фамилии, названия организаций и т. п. Во-вторых, в словарь попала распространенная сниженная лексика. Наконец, многие слова, такие как «авиаперелет», «переориентироваться» и «однодневный», не содержатся в словаре *OpenCorpora*, однако присутствуют в других словарях и не являются неологизмами. Поэтому была произведена дополнительная ручная фильтрация списка несловарных слов авторами статьи, в результате которой были удалены слова, обнаруженные в поисковой системе *Яндекс.Словари*[9] и в Национальном Корпусе Русского Языка (НКРЯ) [Плунгян 2003]. Система *Яндекс.Словари* производит поиск по 10 словарям русского языка, 50 энциклопедиям и 22 словарям иностранных языков. НКРЯ на февраль 2013 года содержал 335 076 текста (364 881 378 словоформ), которые можно разделить на две группы: современные письменные тексты середины XX — начала XXI века и ранние тексты середины XVIII — середины XX века. Мы использовали поиск по всем текстам НКРЯ до 2000 года. Результат данного этапа — список из 168 популярных неологизмов, извлеченных из корпуса текстов социальной сети.

5. Лингвистический анализ словаря неологизмов. Наконец, мы классифицировали полученный список неологизмов по типу заимствования, типу словообразования и модели словообразования (см. таблицу 2, приложение и следующие разделы статьи).

---

[8] Список несловарных слов и неологизмов: https://www.dropbox.com/s/miwknzucui13l60/neo-facebook.xlsx

[9] http://slovari.yandex.ru

Таблица 2. 168 неологизмов Фейсбука, упорядоченных по типу заимствования

Неологизм	Тип заимствования	Тип словообразования	Модель словообразования
сексодром	Англицизм		
айпад, айфон, алерт, байк, бейдж, билдер, блоггинг, брейн, брендинг, вау, виджет, девелопер, демотиватор, дресс, инфо, кавер, караванер, клаб, корпоратив, комент, коммент, коучинг, лайт, лайф, мем, ноут, паблик, перфоманс, плиз, праймериз, принт, продакшн, промо, райдер, ребрендинг, рекрутинг, репост, ретвит, реферал, ритейл, ритейлер, роутер, сиквел, скайп, скрин, сорри, стайл, стор, твитер, твиттер, тизер, трекер, треш, трэш, фейк, форсайт, фреш, фэшн, хайп, холдем, чарт, шутер	Англицизм	Исх 1 корень	
битрейт, бумбокс, геймплей, дабстеп, дедлайн, инфомаркетинг, клипарт, копирайтинг, никнейм, оффлайн, плагин, плейлист, плэйкаст, подкаст, рингтон, стартапер, топфейс, фейсбук, флешмоб, флэшмоб, фолловер, форекс, фрилансер, фэйсбук, хардкор, ютуб, ютюб	Англицизм	Исх 2 корня	
декупаж	Галлицизм		
жжот, капец, мульт, мда, медвед, пипец, ппц, секстиль	Исконное		
госуслуга	Исконное	Композит	ST-ST
единорос	Исконное	Композит	ST-о-ST
всечь	Исконное	Префикс	в-ST
нафиг, нахер, нахрен	Исконное	Префикс	на-ST
предзаказ	Исконное	Префикс	пред-ST
заценить	Исконное	Префикс+суффикс	за-ST-и
офигевать	Исконное	Префикс+суффикс	о-ST-ева
прокремлевский	Исконное	Префикс+суффикс	про-ST-ск
бухарь	Исконное	Суффикс	ST-арь
улыбизм	Исконное	Суффикс	ST-изм
приколист	Исконное	Суффикс	ST-ист
личка, печенька, ржака	Исконное	Суффикс	ST-к
ржачный, улетный	Исконное	Суффикс	ST-н
херня	Исконное	Суффикс	ST-нь
пристройство	Исконное	Суффикс	ST-ств
ржач	Исконное	Суффикс	ST-ч
адчайший	Исконное	Суффикс	ST-ч-айш
днюха	Исконное	Суффикс	ST-юх
вкусняшка	Исконное	Суффикс	ST-яшк
евроинтеграция, инфографика, инфопродукт, телепроект, фотопроект, фотостудия, видеорепортаж	Из заимств. корней	Композит	ST-ST

Неологизм	Тип заимствования	Тип словообразования	Модель словообразования
аудиокнига, вконтакт, мультиварка, нардеп, фотолента, фотоотчет, фотопамять, фотоподборка, фотоприкол, фотошкола	Смешанное	Композит	ST-ST
лохотрон, файлообменник	Смешанное	Композит	ST-о-ST
перепост	Смешанное	Префикс	пере-ST
предстарт	Смешанное	Префикс	пред-ST
забанить, запостить	Смешанное	Префикс+суффикс	за-ST-и
зацикливаться	Смешанное	Префикс+суффикс	за-ST-ива
перепостить	Смешанное	Префикс+суффикс	пере-ST-и
лайкать	Смешанное	Суффикс	ST-а
культурить, постить, твитить	Смешанное	Суффикс	ST-и
анимировать	Смешанное	Суффикс	ST-ирова
аватарка, гифка, флешка	Смешанное	Суффикс	ST-к
реферальный	Смешанное	Суффикс	ST-н
планшетник, цитатник	Смешанное	Суффикс	ST-ник
имхонуть, лайкнуть	Смешанное	Суффикс	ST-ну
брендовый, драйвовый	Смешанное	Суффикс	ST-ов
форумок	Смешанное	Суффикс	ST-ок
суперский	Смешанное	Суффикс	ST-ск
позитивчик	Смешанное	Суффикс	ST-чик
креативщик	Смешанное	Суффикс	ST-щик

3.3. Классификация наиболее частотных несловарных слов

Для понимания языковых изменений важно знать, откуда произошло слово, каким способом оно образовано, каковы его морфологические и синтаксические свойства, и как оно используется. Руководствуясь данным соображением, мы классифицировали полученный список слов по пяти основаниям: *тип заимствования*, *часть речи*, *тип словообразования*, *словообразовательная модель* и *тематика* (см. таблицу в Приложении). Именно эти характеристики мы считаем наиболее показательными в анализе современных заимствований и неологизмов.

По типу заимствования слова из списка разделены на пять классов: исконные, англицизмы, галлицизмы, слова с иноязычными корнями и смешанные слова. Исконными словами считаются те лексемы, которые образованы от общеславянских корней, к англицизмам и галлицизмам причислены новейшие заимствования из английского и французского языков, к словам из заимствованных корней отнесены композиты, составленные из заимствований более раннего времени, тогда как к смешанным словам отнесены слова, состоящие из русскоязычных и иноязычных морфем.

Части речи представлены шестью основными классами: существительное (N), существительное-модификатор (Nmod, подробнее см. раздел 4), прилагательное (Adj), глагол (V), наречие (Adv) и междометие (Interj).

Что касается словообразовательных характеристик, то классификация по типу словообразования соответствует традиционно-грамматическому делению лексем на образованные с помощью суффикса, префикса, префикса и суффикса и словосложения. Также при непроизводных заимствованных словах дополнительно указывается количество корней в языке-источнике (один или два). Словообразовательные модели представляют собой записанную через дефис комбинацию основы слова (ST) и участвующих в словообразовании элементов: префиксов, суффиксов и др.

Наконец, тематикаслов как основание для классификации делит слова данного списка на семантические поля, к которым они относятся. Наиболее многочисленными по составу являются семантические поля «интернет» (*офф-лайн, браузерный*), «оценка» (*суперский, треш/трэш*), «маркетинг» (*реферал, продакшн*) и «мультимедиа» (*фотопроект, плэйкаст*). Следует оговориться, что данное деление на семантические поля условно, поскольку одно и то же слово может принадлежать к разным полям. Поэтому в спорных случаях мы как правило отдавали предпочтение более крупному полю.

4. Полученные результаты и некоторые наблюдения

Предсказуемо большее число неологизмов составляют слова, полностью заимствованные из других языков (93 слова), однако любопытно то, что практически все они, кроме галлицизма *декупаж*, являются словами английского происхождения. Кроме того, есть также семь слов, содержащих иноязычные корни (всего 43), из которых восемь слов составлены из иностранных корней и еще 35 имеют смешанное происхождение, т. е. содержат как иноязычные, так и русские морфемы. Остальные слова (всего 32) являются исконно русскими или калькированными с других языков при помощи русских корней.Наиболее активно пополняемым семантическим полем, по нашим данным, является поле «интернет», представленное практически исключительно англицизмами и образованными от них словами (35 слов). Второе место занимает поле «оценка» (25 слов), поскольку данное семантическое поле является, по всей очевидности, одним из наиболее востребованных, в особенности с появлением социальных сетей и возможностью комментировать выкладываемый в них материал. Стоит также отметить, что к данному полю относится ровно половина исконно русских неологизмов из нашего списка (16 слов). Другие два активно развивающихся поля — «маркетинг» (18 слов) и «техника» (14 слов) — также представлены преимущественно английскими заимствованиями. Наконец, основу еще одного обширно представленного в материале поля «мультимедиа» (15 слов) составляют смешанные слова или композиты из иноязычных корней.

По своей частеречной принадлежности абсолютное большинство слов (123) являются существительными, а также имеется 15 глаголов, 8 прилагательных, 4 междометия и 3 наречно-предикативных слова. Кроме того, имеется особый класс существительных-модификаторов (всего 15 слов), таких как, например, *брейн, лайф* и *фэшн*. Такие единицы в основном англоязычного происхождения в силу незначительной степени адаптированности употребляются

почти всегда в сочетаниях с другими существительными, модифицируя их по принципу английских конструкций типа «stone wall»: *брейн-система, лайф-коуч, фэшн-индустрия* (нередко встречается раздельное написание). Однако среди этих слов встречаются и такие, которые в некоторых случаях могут употребляться самостоятельно, например *байк* и *трэш/треш* (ср. *байк-центр* и *трэш-комедия*). Такие случаи объединены в отдельный промежуточный класс (Nmod/N). В целом такая модель является новой для русского языка.

Что же касается морфологии данных слов, то всего в списке имеется 101 непроизводное и 67 производных слов. При этом стоит отметить, что среди непроизводных слов только восемь являются исконными, тогда как среди производных их число равняется 24. Наиболее распространенным способом словообразования является суффиксация (33). К числу продуктивных суффиксальных моделей можно отнести субстантивную модель с суффиксом -к- (6 слова), в равной степени образующую слова от русских и иноязычных корней, а также глагольную модель с суффиксом -и- (3 слова), имеющую префиксально-суффиксальный коррелят (4 слова). Отдельного внимания заслуживает субстантивная модель с суффиксом -ч-, поскольку является, по всей очевидности, инновацией. К данной модели из списка относится слово *ржач*, однако нам известны и другие случаи, например слова *срач* и *махач*. Других новых моделей на данном материале не зафиксировано. Вторым по распространенности способом является словосложение, представленное, за исключением слов *единорос* и *госуслуга*, словами иноязычного и смешанного происхождения. Из оставшихся префиксального и префиксально-суффиксального способа словообразования (по 7 слов) наиболее продуктивными являются, соответственно, модель с префиксом на- и с префиксом за- и суффиксом -и- (по 3 слова).

Если обобщить изложенные выше данные, то в настоящее время наблюдается активный поток заимствований из английского языка, окончательно закрепившего за собой статус языка международного общения, о чем свидетельствует абсолютное превосходство англицизмов над всеми остальными неологизмами из нашего списка. Практически все остальные слова возникают в результате комбинаций уже существующих в языке исконных и иноязычных морфем. Основными активно обновляющимися в лексическом плане сферами являются мультимедиа- и интернет-технологии, а также торговля, что свидетельствует о существенной роли данных областей в современной жизни. Что характерно, обновление данных полей, кроме поля «мультимедиа», происходит практически исключительно за счет англицизмов. Неологизмы с исконными корнями же возникают прежде всего в оценочной сфере в силу растущей потребности выражать собственное мнение о материалах, выкладываемых в социальные сети. Большинство как исконных, так и заимствованных неологизмов являются существительными, поскольку данная часть речи является наиболее подходящей для описания новых для языка реалий. Заимствованные неологизмы, являясь по большей части существительными, преимущественно непроизводные, и только часть слов, адаптированных в системе русского языка, содержит словообразовательные морфемы. Исконные слова же, закономерным образом, почти все являются производными, поскольку единственным, помимо заимствования способом образования неологизмов является деривация.

5. Заключение

В результате нашей работы полуавтоматическим способом был получен словарь из 168 неологизмов русскоязычного сегмента социальной сети Фейсбук. Найденные неологизмы представляют собой наиболее частотную часть списка и активно используются и в других социальных сетях и платформах в интернете. Таким образом, мы создали словарную базу, которая может быть использована для различных исследований, касающихся как возникновения неологизмов, так и процесса заимствования и адаптации заимствованных слов в языке.

Кроме того, мы предложили классификацию полученного материала по типу заимствования, тематике, части речи, типу словообразования и модели словообразования. Данная классификация позволяет получить первоначальное представление о специфике морфологии и синтаксиса новейших лексических единиц языка и происходящих в наше время лексических процессах, дает возможность обратить внимание на актуальные сферы использования современной лексики и сравнить свойства исконных и заимствованных слов.

Наконец, мы надеемся, что результаты наших исследований помогут улучшить понимание принципов автоматической обработки и анализа слов, отсутствующих в словарях, что должно улучшить состояние технологий обработки текста.

Литература

1. *Брейтер М. А.* (1997), Англицизмы в русском языке: история и перспективы: Пособие для иностранных студентов-русистов. // Владивосток, Диалог-МГУ.
2. *Дьяков А. И.* (2003), Причины интенсивного заимствования англицизмов в современном русском языке. // Язык и культура. — Новосибирск. — С. 35–43.
3. *Зализняк, А. А.* (1977). Грамматический словарь русского языка: словоизменение: около 100 000 слов. // Изд-во «Русский язык».
4. *Крысин Л. П.* (1968) Иноязычные слова в современном русском языке — М.: Наука
5. *Плунгян, В. А.* (2003) «Зачем нужен Национальный корпус русского языка?» Национальный корпус русского языка: 6–21.
6. *Маринова Е. В.* (2013) Иноязычная лексика современного русского языка: учеб. пособие, 2-е изд., М. : ФЛИНТА
7. *Capuz Juan Gómez*. (1997), Towards a Typological Classification of Linguistic Borrowing (Illustrated with Anglicisms in Romance Languages) // Revista Alicantina de Estudios Ingleses 10: 81–94
8. *Catanese S., De Meo P., Ferrara E., Fiumara E., and Provetti A.* (2012). Extraction and analysis of Facebook friendship relations. // In Computational Social Networks, pages 291–324. Springer.
9. *Cougnon, L.-A., and François T.* (2010), Quelques contributions des statistiques à l'analyse sociolinguistique d'un corpus de SMS. // Actes du colloque JADT 2010. Vol. 1. 2010.

10. *Estopa Rosa* (2011), Neologisms at the Boundaries of Prefixation, Composition and Syntagmatic Composition in Catalan: Controversial and Open Questions // Organon 25.50.
11. Facebook: Key Facts. Режим доступа: http://newsroom.fb.com/Key-Facts
12. *Fischer R.* (1950), Lexical change in present day English: a corpus-based study of the motivation, institutionalization and productivity of creative neologisms.
13. *Hall, T. A. & Hamann, S.* (2003). Loanword Nativization in German. // Zeitschrift für Sprachwissenschaft 22, 56–85.
14. *Haugen E.* (1950), The analysis of linguistic borrowing. // Language: 210–231
15. *Heinemann, S.* (2003). Hai letto il mio [i'mɛil]? Anmerkungen zur lautlichen Adaption von Anglizismen im Italienischen. // Romanische Forschungen 115, Internet World Stats. Miniwatts Marketing Group. Режим доступа: http://www.internetworldstats.com/stats4.htm
16. *Jennifer L.* (2007), Smith. Source similarity in loanword adaptation: Correspondence Theory and the posited source-language representation. // Phonological Argumentation: Essays on Evidence and Motivation, London: Equinox
17. *Khurshid Ahmad* (1991), Neologisms, Nonces and Word Formation // Language Change: Decay or Evolution. Cambridge Univ. Press.
18. *LaCharité, D. & Paradis, C.* (2005). Category Preservation and Proximity versus Phonetic Approximation in Loanword Adaptation. // Linguistic Inquiry 36/2, 223–258.
19. *Lui M. and Baldwin T.* (2012). langid. py: An off-the-shelf language identification tool. // In Proceedings of the ACL 2012 System Demonstrations, pages 25–30. Association for Computational Linguistics.
20. *Peperkamp, S. & Dupoux, E.* (2003). Reinterpreting loanword adaptations: The role of Perception // Proceedings of the 15th International Congress of Phonetic Sciences 2003, 367–370.
21. *Volland, B.* (1986). Französische Entlehnungen im Deutschen. Transferenz und Integration auf phonologischer, graphematischer, morphologischer und lexikalischsemantischer Ebene. Tübingen: Niemeyer.
22. *Winter-Froemel E.* (2008), Studying loanwords and loanword integration: Two criteria of conformity // Newcastle Working Papers in Linguistics 14: 1

Приложение 1. Словарь неологизмов русскоязычного Фейсбука

#	Слово	Часть речи	Тематика	Тип заимствования	Тип словообразования	Модель словообразования	Частота
1	лайкать	V	интернет	Смешанное	Суффикс	ST-а	34222
2	вконтакт	N	интернет	Смешанное	Композит	ST-ST	1134119
3	перепост	N	интернет	Смешанное	Префикс	пере-ST	426927
4	постить	V	интернет	Смешанное	Суффикс	ST-и	61998
5	твитить	V	интернет	Смешанное	Суффикс	ST-и	100988
6	файло-обменник	N	интернет	Смешанное	Композит	ST-о-ST	27638
7	аватарка	N	интернет	Смешанное	Суффикс	ST-к	22432
8	забанить	V	интернет	Смешанное	Префикс+суффикс	за-ST-и	18400

#	Слово	Часть речи	Тематика	Тип заимствования	Тип словообразования	Модель словообразования	Частота
9	запостить	V	интернет	Смешанное	Префикс+суффикс	за-ST-и	21971
10	перепостить	V	интернет	Смешанное	Префикс+суффикс	пере-ST-и	212364
11	цитатник	N	интернет	Смешанное	Суффикс	ST-ник	59566
12	лайкнуть	V	интернет	Смешанное	Суффикс	ST-ну	68457
13	имхонуть	V	интернет	Смешанное	Суффикс	ST-ну	17395
14	форумок	N	интернет	Смешанное	Суффикс	ST-ок	26937
15	личка	N	интернет	Исконное	Суффикс	ST-к	247497
16	блоггинг	N	интернет	Англицизм			22332
17	комент	N	интернет	Англицизм			30312
18	коммент	N	интернет	Англицизм			117196
19	никнейм	N	интернет	Англицизм			17702
20	оффлайн	N	интернет	Англицизм			21614
21	паблик	N	интернет	Англицизм			48587
22	репост	N	интернет	Англицизм			144676
23	ретвит	N	интернет	Англицизм			29336
24	скайп	N	интернет	Англицизм			426877
25	твитер	N	интернет	Англицизм			22751
26	твиттер	N	интернет	Англицизм			578503
27	топфейс	N	интернет	Англицизм			161399
28	трекер	N	интернет	Англицизм			45888
29	фейсбук	N	интернет	Англицизм			377884
30	фолловер	N	интернет	Англицизм			22124
31	фэйсбук	N	интернет	Англицизм			62267
32	ютуб	N	интернет	Англицизм			83074
33	ютюб	N	интернет	Англицизм			24043
34	битрейт	N	интернет	Англицизм			58676
35	инфографика	N	интернет	Англицизм			91729
36	культурить	V	культура	Смешанное	Суффикс	ST-и	24073
37	секстиль	N	культура	Исконное			117415
38	дабстеп	N	культура	Англицизм			18255
39	демотиватор	N	культура	Англицизм			22432
40	кавер	N	культура	Англицизм			72386
41	клипарт	N	культура	Англицизм			23663
42	мем	N	культура	Англицизм			38912
43	перфоманс	N	культура	Англицизм			23905
44	плэйкаст	N	культура	Англицизм			367575
45	сиквел	N	культура	Англицизм			17685
46	флешмоб	N	культура	Англицизм			84966
47	флэшмоб	N	культура	Англицизм			34780
48	предстарт	N	маркетинг	Смешанное	Префикс	пред-ST	23441
49	реферальный	Adj	маркетинг	Смешанное	Суффикс	ST-н	93330
50	предзаказ	N	маркетинг	Исконное	Префикс	пред-ST	22915
51	инфопродукт	N	маркетинг	Из заимств. корней	Композит	ST-ST	17824
52	алерт	N	маркетинг	Англицизм			19949
53	брендинг	N	маркетинг	Англицизм			33457
54	инфомаркетинг	N	маркетинг	Англицизм			17928
55	продакшн	N	маркетинг	Англицизм			33134
56	промо	Nmod	маркетинг	Англицизм			134124
57	ребрендинг	N	маркетинг	Англицизм			19333
58	реферал	N	маркетинг	Англицизм			103650
59	ритейл	N	маркетинг	Англицизм			21971
60	ритейлер	N	маркетинг	Англицизм			23151
61	стартапер	N	маркетинг	Англицизм			21100
62	стор	Nmod	маркетинг	Англицизм			36337
63	тизер	N	маркетинг	Англицизм			49417
64	форекс	N	маркетинг	Англицизм			186372
65	хайп	N	маркетинг	Англицизм			31586
66	анимировать	V	мультимедиа	Смешанное	Суффикс	ST-ирова	33222

#	Слово	Часть речи	Тематика	Тип заимствования	Тип словообразования	Модель словообразования	Частота
67	гифка	N	мультимедиа	Смешанное	Суффикс	ST-к	28293
68	фотоотчет	N	мультимедиа	Смешанное	Композит	ST-ST	129303
69	фотопамять	N	мультимедиа	Смешанное	Композит	ST-ST	67254
70	фотоподборка	N	мультимедиа	Смешанное	Композит	ST-ST	25999
71	фотоприкол	N	мультимедиа	Смешанное	Композит	ST-ST	85613
72	мульт	N	мультимедиа	Исконное			76768
73	видеорепортаж	N	мультимедиа	Из заимств. корней	Композит	ST-ST	17665
74	телепроект	N	мультимедиа	Из заимств. корней	Композит	ST-ST	26985
75	фотолента	N	мультимедиа	Из заимств. корней	Композит	ST-ST	18920
76	фотопроект	N	мультимедиа	Из заимств. корней	Композит	ST-ST	45528
77	фотостудия	N	мультимедиа	Из заимств. корней	Композит	ST-ST	31041
78	фотошкола	N	мультимедиа	Из заимств. корней	Композит	ST-ST	17942
79	лайф	Nmod	мультимедиа	Англицизм			28368
80	скрин	N	мультимедиа	Англицизм			23283
81	брендовый	Adj	одежда	Смешанное	Суффикс	ST-ов	26893
82	бейдж	N	одежда	Англицизм			17981
83	дресс	Nmod	одежда	Англицизм			32026
84	принт	N	одежда	Англицизм			81120
85	фэшн	Nmod	одежда	Англицизм			32895
86	лохотрон	N	оценка	Смешанное	Композит	ST-о-ST	30760
87	драйвовый	Adj	оценка	Смешанное	Суффикс	ST-ов	24135
88	суперский	Adj	оценка	Смешанное	Суффикс	ST-ск	19754
89	позитивчик	N	оценка	Смешанное	Суффикс	ST-чик	22556
90	нафиг	Adv/Pred	оценка	Исконное	Префикс	на-ST	39668
91	приколист	N	оценка	Исконное	Суффикс	ST-ист	69330
92	нахер	Adv/Pred	оценка	Исконное	Префикс	на-ST	17553
93	нахрен	Adv/Pred	оценка	Исконное	Префикс	на-ST	22916
94	ржака	N	оценка	Исконное	Суффикс	ST-к	43288
95	ржачный	Adj	оценка	Исконное	Суффикс	ST-н	22716
96	жжот	V	оценка	Исконное			31450
97	капец	N	оценка	Исконное			27650
98	заценить	V	оценка	Исконное	Префикс+суффикс	за-ST-и	200777
99	пипец	N	оценка	Исконное			51314
100	улетный	Adj	оценка	Исконное	Суффикс	ST-н	19758
101	ппц	N	оценка	Исконное			29883
102	херня	N	оценка	Исконное	Суффикс	ST-нь	28836
103	офигевать	V	оценка	Исконное	Префикс+суффикс	о-ST-ева	26990
104	ржач	N	оценка	Исконное	Суффикс	ST-ч	32035
105	адчайший	Adj	оценка	Исконное	Суффикс	ST-ч-айш	22432
106	премиум	Nmod	оценка	Англицизм			58276
107	треш	Nmod/N	оценка	Англицизм			17421
108	трэш	Nmod/N	оценка	Англицизм			27304
109	фейк	N	оценка	Англицизм			24691
110	хардкор	N	оценка	Англицизм			20181
111	вкусняшка	N	питание	Исконное	Суффикс	ST-яшк	30777
112	фреш	Nmod/N	питание	Англицизм			18830
113	нардеп	N	политика	Смешанное	Композит	ST-ST	44061
114	госуслуга	N	политика	Исконное	Композит	ST-ST	17942
115	единорос	N	политика	Исконное	Композит	ST-о-ST	17405
116	прокремлевский	Adj	политика	Исконное	Префикс+суффикс	про-ST-ск	18001
117	евроинтеграция	N	политика	Из заимств. корней	Композит	ST-ST	16984
118	праймериз	N	политика	Англицизм			23514
119	зацикливаться	V	психология	Смешанное	Префикс+суффикс	за-ST-ива	17096

#	Слово	Часть речи	Тематика	Тип заимствования	Тип словообразования	Модель словообразования	Частота
120	улыбизм	N	психология	Исконное	Суффикс	ST-изм	19457
121	коучинг	N	психология	Англицизм			100986
122	девелопер	N	работа	Англицизм			22447
123	дедлайн	N	работа	Англицизм			19996
124	корпоратив	N	работа	Англицизм			39322
125	рекрутинг	N	работа	Англицизм			17694
126	фрилансер	N	работа	Англицизм			39902
127	печенька	N	разное	Исконное	Суффикс	ST-к	27218
128	всечь	V	разное	Исконное	Префикс	в-ST	24566
129	мда	Interj	разное	Исконное			51107
130	медвед	N	разное	Исконное			19568
131	пристройство	N	разное	Исконное	Суффикс	ST-ств	24515
132	декупаж	N	разное	Галлицизм			29524
133	инфо	N	разное	Англицизм			144808
134	брейн	Nmod	разное	Англицизм			104056
135	вау	Interj	разное	Англицизм			47633
136	копирайтинг	N	разное	Англицизм			53164
137	лайт	Nmod/N	разное	Англицизм			21867
138	плиз	Interj	разное	Англицизм			60359
139	райдер	N	разное	Англицизм			26129
140	сексодром	N	разное	Англицизм			27398
141	сорри	Interj	разное	Англицизм			27449
142	стайл	Nmod/N	разное	Англицизм			23461
143	форсайт	Nmod	разное	Англицизм			22689
144	чарт	N	разное	Англицизм			29022
145	байк	Nmod/N	спорт и досуг	Англицизм			24999
146	билдер	N	спорт и досуг	Англицизм			31338
147	геймплей	N	спорт и досуг	Англицизм			19222
148	караванер	N	спорт и досуг	Англицизм			44569
149	клаб	Nmod	спорт и досуг	Англицизм			37051
150	холдем	N	спорт и досуг	Англицизм			43273
151	шутер	N	спорт и досуг	Англицизм			28704
152	флешка	N	техника	Смешанное	Суффикс	ST-к	56179
153	планшетник	N	техника	Смешанное	Суффикс	ST-ник	25473
154	аудиокнига	N	техника	Смешанное	Композит	ST-ST	73058
155	мультиварка	N	техника	Смешанное	Композит	ST-ST	18704
156	айпад	N	техника	Англицизм			38430
157	айфон	N	техника	Англицизм			111594
158	бумбокс	N	техника	Англицизм			47473
159	виджет	N	техника	Англицизм			25507
160	ноут	N	техника	Англицизм			30447
161	плагин	N	техника	Англицизм			79222
162	плейлист	N	техника	Англицизм			256455
163	подкаст	N	техника	Англицизм			68787
164	рингтон	N	техника	Англицизм			20499
165	роутер	N	техника	Англицизм			18340
166	креативщик	N	человек	Смешанное	Суффикс	ST-щик	19408
167	бухарь	N	человек	Исконное	Суффикс	ST-арь	23749
168	днюха	N	человек	Исконное	Суффикс	ST-юх	18235

ПРИМЕНЕНИЕ МОДЕЛИ УСЛОВНЫХ СЛУЧАЙНЫХ ПОЛЕЙ ДЛЯ ОПРЕДЕЛЕНИЯ МОРФОЛОГИЧЕСКИХ ХАРАКТЕРИСТИК СЛОВ В РУССКОЯЗЫЧНЫХ ТЕКСТАХ

Музычка С. А. (s.muzychka@samsung.com)

Московский государственный университет
им. М. В. Ломоносова, Москва, Россия;
ООО «Исследовательский центр Самсунг», Москва, Россия

Романенко А. А. (a.romanenko@samsung.com)

Московский физико-технический институт
(государственный университет), Москва, Россия;
ООО «Исследовательский центр Самсунг», Москва, Россия

Пионтковская И. И. (p.irina@samsung.com)

ООО «Исследовательский центр Самсунг», Москва, Россия

В статье рассматривается проблема снятия морфологической омонимии для русского языка с помощью статистических методов, а именно аппарата условных случайных полей (англ. *Conditional Random Fields,* CRF). Предлагается модифицированная модель CRF, дающая результаты, соответствующие state-of-the-art.
Также рассматривается применение CRF для нормализации цифровой записи числительных. Приводятся результаты вычислительного эксперимента.

Ключевые слова: снятие морфологической омонимии, условное случайное поле, CRF, нормализация текста, NLP

CONDITIONAL RANDOM FIELD FOR MORPHOLOGICAL DISAMBIGUATION IN RUSSIAN

Muzychka S. A. (s.muzychka@samsung.com)
Lomonosov Moscow State University, Moscow, Russia;
Samsung R&D Institute Rus, Moscow, Russia

Romanenko A. A. (a.romanenko@samsung.com)
Moscow Institute of Physics and Technology, Moscow, Russia;
Samsung R&D Institute Rus, Moscow, Russia

Piontkovskaja I. I. (p.irina@samsung.com)
Samsung R&D Institute Rus, Moscow, Russia

We consider the problem of morphological disambiguation in Russian using statistical methods; specifically, we apply conditional random field (CRF). We propose a new modified model of linear chain CRF, which demonstrates results close to the state-of-the-art. We also propose a new statistical approach to text normalization problem using CRF. Namely, we solve the problem of normalization of numerals written as digits. Our approach allows for the consideration of both cardinal and ordinal numbers.
In order to train and test our models we used Russian text corpora. For morphological disambiguation, we used data from OpenCorpora and the SynTagRus linguistic corpus. For number normalization we used the Russian National Corpora (RusCorpora).
A brief overview of the CRF model is given, followed by a detailed description of the applied algorithm, assumptions on the training and test set, and a description of features for each particular issue.

Key words: morphological disambiguation, conditional random field, text normalization, NLP

1. Introduction

One of the key problems in text processing is morphological disambiguation. The results of this analysis can be applied to another natural language processing (NLP) problems: extracting named entities, syntactic parsing, sentiment analysis, etc.

While it is considered this problem to be solved for English language, there are some difficulties for languages with rich morphology, in particular for russian language. To solve the problem for Russian language both rule-based and statistical approaches are applied. The main obstacles in the solution of the problem are, firstly, the lack of a single common morphological tagset and, secondly, absence of common training and test corpora for verification of implemented algorithms.

CRF algorithm presents state-of-the-art results in many NLP problems related to sequence labeling. In this articles we consider an application of CRF algorithm for the solution of two problems: morphology disambiguation and number normalization. To solve this problem we propose modified CRF models that enable to reduce learning time.

The rest of the paper is structured as follows. Section 2 introduces basic concepts of CRF models, section 3 describes its application for morphological disambiguation, and section 4 presents an algorithm for number normalization. Each section contains description of used data and results.

2. Conditional Random Fields

This section describes basic concepts of CRF models introdused in [9].

2.1. General CRF Model

Definition 1. Let $G = (V, E)$ be an undirected graph. The set of random variables $\{\xi_v\}, v \in V$ form a Markov random field (MRF) with respect to G if they satisfy the local Markov properties:

1. **Pairwise Markov property**: any two non-adjacent variables ξ_x and ξ_y are conditionally independent given all other variables $\{\xi_v\}_{v \in V \setminus \{x,y\}}$.
2. **Local Markov property:** each variable ξ_x is conditionally independent of all other variables given its neighbors $\{\xi_v\}_{v:\{x,v\} \in E}$.
3. **Global Markov property:** any two subsets of variables $\{\xi_v\}_{v \in A}$, $\{\xi_v\}_{v \in B}$, $A, B \subset V$, $A \cap B = \emptyset$, are conditionally independent given a separating subset ($S \subset V$ the set of nodes is called separating for two non-intersecting subsets if each path from the first subset to the second passes through it).

Definition 2. The maximal clique of the graph G is called any maximal fully connected subgraph of G.

Theorem (*Hammersely-Clifford*) *The set of random variables $\xi = \{\xi_v\}, v \in V$ is a MRF corrsesponding to the graph G if and only if its distribution $p(\xi)$ is factorized by cliques of the graph, that is*

$$p(\xi) = \frac{1}{Z} \prod_{c \in C} \Psi_c(\xi),$$

where C is the set of all maximal cliques of G, Ψ_c are some functions depending on $\xi = \{\xi_v\}, v \in c$ only, and Z is a normalization factor called partition function.

Definition 3. CRF is an MRF such that the set of its nodes is decomposed into two noninteracting subsets $V = X \cup Y$ where X and Y are the sets of observed and hidden variables correspondingly.

Everywhere below we use the following notations $x = \{\xi_v : v \in X\}$ and $y = \{\xi_v : v \in Y\}$. Also we suppose that the random variables from x and y belong to some arbitrary discrete spaces X and Y correspondingly.

The inference problem is to predict the optimal values of y, given the observations x. According to Hammersely-Clifford theorem we should optimize

$$p(y \mid x) = \frac{1}{Z(x)} \prod_{c \in C} \Psi_c(x, y), \text{ where } Z(x) = \sum_{y'} \prod_{c \in C} \Psi_c(x, y') \text{ is a partition function.}$$

Usually Ψ_c are chosen as an exponent of the linear combination of some features with coefficients that should be determined during training. As a rule these coefficients depend on the structure of the clique only and given training set $Tr = \{(x^j, y^j)\}$ are fitted by maximizing the log-likelihood probability

$$\ln p(Tr) = \sum_j \ln p(y^j \mid x^j).$$

2.2. Linear Chain CRF

The structure of the model is depicted on Figure 1.

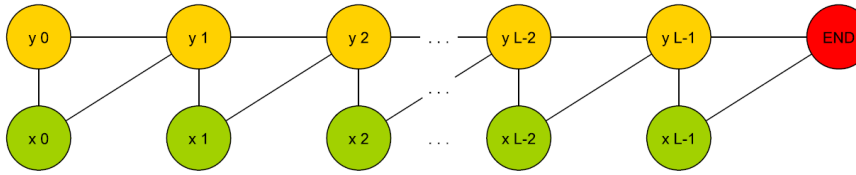


Fig. 1. Linear chain CRF

The nodes $y_0, y_1, ..., y_{L-1}$ correspond to our hidden parameters y, and $x_0, x_1, ..., x_{L-1}$ correspond to the observations x. The last node END is terminal and its value without loss of generalty can be set to any fixed element which we for simplicity denote by 'end'. Every clique in the graph consists of two consequitive hidden nodes y_k, y_{k+1} and one observable node x_k. Consequently, the functions Ψ_c inroduced above have the form $\Psi(x, y', y'')$. The observations x_k are usually represented as a vector of binary features

$$x_k = (f_k^1, f_k^2, ..., f_k^F), f_k^i \in \{0,1\}.$$

Finally, $\Psi(x, y', y'')$ are chosen in the form

$$\Psi(x, y', y'') = \exp\left(\sum_{i=1}^{F} \lambda(i, y', y'') f_k^i + b(y', y'') \right)$$

where $\lambda(i, y', y'')$ and $b(y', y'')$ are parameters of the model that should be tuned during training.

In order to optimize maximum likelihood function any gradient descent method can be used.

3. Morphological Disambiguation

3.1. Previous Results

Currently, there are a large number of papers devoted to the definition of parts of speech (POS-tagging). For example, the papers [1,5] describe classifiers with accuracy 95–97%. We consider the problem of full morphological disambiguation (POS+MORPH), that is for each token in the sentence we should assign the corresponding morphological labels: part of speech, gender, animacy, etc.

Note that is not always possible to quantitatively compare two different systems of analysis of the Russian language since often they use different tagsets. Also sometimes punctuation tokens are used for calculating accuracy of the algorithm. Also it is important which morphological dictionary is used: some systems use external resources (for example, Zaliznyak dictionary); the others extract dictionary from training data [5].

In paper [6], evaluation of morphological analysis on full tagset is done. Their tagset consists of 829 tags. The presented result is 94,46%. But, as far as it can be understood from the article, there is no unknown (out-of-vocabulary) words in their test set.

In [5], the result 95.25% is performed on universal MTE tagset.

Finally, there are corresponding disambiguation algorithms with high accuracy of classification for such morphologically rich languages as Czech, Hungarian, etc. The following table represents results for them reported in [2]:

Language	**arabic**	**czech**	**spanish**	**german**	**Hungarian**
Number of morphological labels	516	1811	303	681	1071
Accuarcy	90.32	92.94	97.93	88.58	96.34

3.2. Application of CRF for Morphological Disambiguation

For most languages, the standard CRF model is well suited to solve this problem and provides a practically significant results. However, for such morphologically rich languages like Russian and Czech, where the total number of morphological markers of several hundred, the direct application of the linear model of CRF may cause technical problems.

Firstly, the complexity of gradient computation in the model described above is a quadratic function of the total number of hidden states of the model, and therefore the algorithm converges too slowly. Secondly, the total number of parameters of the models is quadratic in the number of hidden states Во-вторых, количество свободных параметров модели растет квадратично с ростом числа скрытых параметров, which, in turn, increases the complexity of the model. Finally, there is a possibility of attributing any morphological tags to any token in a sentence (for example, there is a possibility to assign label «adjective» to the token «кошка»). At the same time, as a rule, for each token, we have to choose from 3–4 parsing options. To overcome the described difficulties, we consider the classic model of linear CRF in a modified form.

3.3. Description of the Data

To train and test our model, we use the Syntagrus corpus [3]. The size of this corpora is significantly lower than the corresponding analogs – OpenCorpora and RusCorpora (National corpora of Russian language) (total number of sentences is 53439, total number of tokens is 774368). Nevertheless, Syntagrus has certain advantages. Firstly, it has uniqely defined labels. For comparison, OpenCorpora provides only possible morphological labels for each token and therefore can not be used for learning. Secondly, Syntagrus provides marking for syntax trees, which makes it possible to apply the developed algorithm in subsequent parsing.

3.4. Assumptions

During training and testing we made the following simplifications:

1. Labels that contained НЕСТАНД, МЕТА and НЕПРАВ were replaced by UNKNOWN.
2. All tokens with latinic symbols were marked as NID.
3. The difference between the comparative adjectives and comparative adverbs is not considered.
4. Morphological labels СТРАД, СЛ, ПРЕВ are not considered.

To conclude the total number of labels is 353.

3.5. Morphological Dictionary

Since Syntagrus is a rather small corpora we can't use data only from it in order to compose an acceptable dictionary. Consequently, we use dictionary from OpenCorpora project (based on Zalizn'yak grammar dictionary), and then converts marks into SynTagRus format for convinience.

3.6. Description of the Algorithm

The structure of our model is depicted on the following figure

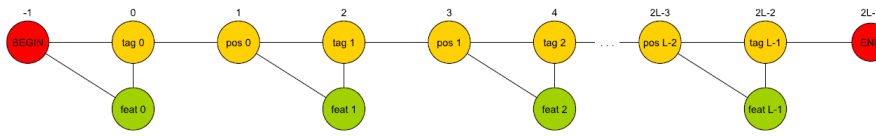


Fig. 2. CRF model, used for morphological disambiguation

The set of hidden elements is decomposed into union of two non-intersecting subsets $Y = labels \cup pos$, where *labels* is the set of all morphological labels, and *pos* is the

set of all possible parts of speech. On Figure 2 the nodes $y_{2k} = l_k \in labels$ corresponds to the full morphological label of the *k*-th token of the sentence and $y_{2k+1} = p_k \in pos$ to its POS. The nodes $y_{-1} = begin$ and $y_L = end$ are initial and terminal nodes. Their content can be set to any arbitrary values and we without loss of generality prefer to denote them 'begin' and 'end' correspondingly. We see that the graph contains two types of cliques. The cliques of the first type consist of one *label*-node y_{2k}, one *pos*-node y_{2k-1} and one feature node x_k. Therefore this type of cliques can be described by function $\Psi(x, y', y'')$. The cliques of the second type correspond to the transition from *label*-node y_{2k} and one *pos*-node y_{2k+1} and are described by function $\Psi(y', y'')$. We assume that this transition is deterministic i.e. we set

$$\Psi(y', y'') = I\{POS\ of\ y' and\ y'' coincide\}.$$

Note that the imposed restrictions allows to solve solve the two first problems described above, namely, the number of free parameters of the model is now proportionally to $|labels| \cdot |pos|$ that is significantly slower than $|labels|^2$. Finally for each node y_k we store a list of possible morphological labels $list_k$. During inference and gradient descent computations we can take into considerations only that label sequences that satisfy the imposed restrictions. This simultaneously helps to prevent very serious mistakes in the prediction of hidden labels, and reduce training time.

3.7. Features

The basic features used in experiments are possible morphological labels derived from OpenCorpora dictionary. Also we the most used tokens (conjunctions, particles, etc.) and token endings were added to the feature set.

3.8. Results

We used first 45,000 sentences for training, and the rest 8439 sentences for testing. The number of tokens in the test set is 121 968. The accuracy of the algorithm is 91,06% on the test set. This result is worth than the similar results from papers [5,6]. But distinctive feature of our algorithm is that it doesn't use lexical information at all. In [3] it is shown that lexical features could improve overall accuracy up to 9%. Our experiments show that even without lexical information statistical approach can be applied to the problem of morphological disambiguation and it performs satisfactory results.

The following table shows the distribution of errors on the part of speech.

	a	adv	com	conj	intj	nid	num	part	pr	S	unknown	v
a	1494	204	5	2	0	36	94	5	3	283	5	140
adv	198	117	1	119	1	20	20	72	4	165	3	15
com	1	0	0	0	0	6	0	0	0	3	0	0

	a	adv	com	conj	intj	nid	num	part	pr	S	unknown	v
conj	8	144	0	0	0	0	0	176	0	135	0	0
intj	0	0	0	0	0	5	0	0	2	0	0	0
nid	5	4	0	0	0	0	0	0	0	67	0	2
num	52	1	0	0	0	8	131	0	0	17	0	0
part	42	38	0	263	2	2	0	0	0	92	0	13
pr	6	74	1	0	0	2	0	0	0	28	0	0
s	304	169	7	72	5	292	38	22	10	4645	23	84
unknown	3	4	0	3	0	2	0	3	0	22	0	0
v	203	28	1	7	2	13	8	13	3	69	0	507

Here rows of the table correspond to the real parts of speech, and columns to the predicted. The diagonal elements contain the number of errors associated with incorrect additional morphological labels (case, gender, etc.) in particular, we see that the greatest number of errors occur within the group of nouns (S) and adjectives (A). The latter is due to the fact that for the distinction of the nominative and accusative cases it is necessary to consider further links in a sentence that is not guaranteed by a linear model CRF. Finally, the accuracy of the classifier on the reduced tagset containing only parts of speech is 96,7% that is consistent with the existing analogues.

4. Normalization of Russian Numerals

The one common task, which appears in Text-To-Speech System development process, is normalization of non-standard words such as abbreviations, acronyms or numerals written in digits [4, 7]. TTS system should be correctly pronoun non-standard phrase, so it should can decode word and set them to proper grammar form.

In this section we give an example of application of linear-chain CRF to detection of grammar form of Russian numerals written in digits. Previous work [8] describes variant of solution of this task based only on frequency features, not on grammar features of contexts of numerals. Also solution described in [8] makes one strict assumption: all numerals written in digits are cardinal that often breaks down in practice. We avoid this assumption and propose to use grammar features of words from context of numerals.

4.1. Dataset and Feature Generation

We used subset of National Russian corpora. Phrases containing numerals with grammar features were selected and all numerals were converted to digit form.

Features that we generate for tokens can divide into 5 parts.

1. Grammar features of words (not numerals). In practice we use disambiguation tool described above.
2. Features indicated that word is specific for cardinal or ordinal numerals. For instance, names of currency usually take place with cardinal numerals.

3. Features indicate prepositions, because they help determine case of nominals.
4. Features that indicate spelling characteristics of numerals: length in symbols, terminal digit, etc.
5. Features from group 1–4 for neighbor tokens

4.2. CRF Model Details

We use modified model of linear-chain CRF in the same manner as in morphological disambiguation algorithm. This model is shown below on figure 3.

We maximize probability of labels sequence "POS of previous token, type and grammar form of numerals, POS of next token" when we have observed sequence "features of previous token $Feat_{prev}$, features of numerals $Feat_{num}$, features of next token $Feat_{next}$. In addition we present type and grammar form of numerals like sequence of five labels: TYPE (cardinal or ordinal), CASE, GEN (masculine, feminine, neuter or unknown), SNGL (singular, plural or unknown) and ANIM (animacy, inanimacy or unknown). We include in features of token features of 7 right and 7 left neighbor tokens.

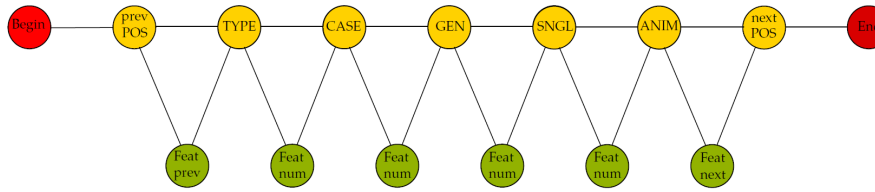


Fig. 3. Model CRF used for detection type and grammar form of numerals

4.3. Experiments

As noticed above we use for training and testing algorithm subset of National Russian corpora (10268 phrases with numerals). We split all data set on training (8251 phrases) and testing (2017) parts.

Accuracy of result model on test set is Also in table below we show result of detection type and grammar form of numerals averaged by category of labels. We evaluate P precision, R recall and F_1-measure as quality measures.

Quality measure	TYPE, %	CASE, %	GEN, %	SNGL, %	ANIM, %
P	97.21	91.33	89.77	82.39	87.66
R	97.21	92.93	90.74	85.97	95.05
F_1	97.21	92.10	90.24	84.05	91.11

Moreover, we evaluate quality of model with 5-fold cross-validation procedure. Result of evaluation is so applying model has high generalization ability.

4.4. Remarks and Result Analysis

Type and grammar form of numerals which are predicted by model described above allow convert numerals to form of words with help of finite-state automaton. This finite-state automaton and model CRF described above give a system of numerals normalization.

It is needed to be noticed that actual accuracy of system should be higher. Firstly, numerals with different grammar features in form of words sometimes match together, and CRF model mistakes usually in these cases. For example, model confuses on nominative and accusative cases. Secondly, when system detects definite gender, animacy or number instead of label unknown it is not mistake. Second note considered accuracy of algorithm rises to 94.53%.

Nevertheless, using in text numerals in form of digit or in form of words are depends on many conditions: kind of text, narrative style of author, etc. But we use as training data phrases with numerals largely in form of words. So, the issue about data quality and quality of model remains open.

5. Conclusions

In article we give some examples of applications of conditional random fields to important tasks of natural language processing. The accuracy of disambiguating algorithm based on CRF is 91.06%, for task of normalization of numerals accuracy is 94.53%. As result of paper we mark up that quality of machine learning approach to NLP tasks for Russian is at a level of quality of rule-based analogues but statistic approach needs less human resources.

References

1. *Antonova A. Ju., Solovyev A. N.* (2013), Conditional random field models for the processing of russian [Ispol'zovanie metoda uslovnyh sluchajnyh polej dlja obrabotki tekstov na russkom jazyke], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2013" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2013"], Bekasovo, pp. 39–52.
2. *Muller T., Schmid H., Schutze H.* (2013), Efficient Higher-Order CRFs for Morphological Tagging, In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, p. 322–332.
3. *Nivre J., Boguslavsky M., Iomdin L.* (2008), Parsing the SynTagRus treebank of Russian, Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pp. 641–648.
4. *Olinsky C., Black A. W.* (2000), Non-standard word and homograph resolution for asian language text analysis, In proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH), ISCA, pp. 733–736

5. *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", Bekasovo, pp. 591–605.
6. *Sokirko, A. V., Toldova, S. Y.* (2004). Comparing a stochastic tagger based on Hidden Markov Model with a rule-based tagger for Russian. Proc. of Corpus Linguistics–2004, Saint-Petersburg. [In Russian].
7. *Sproat R., Black A. W., Chen S. F. and Kumar Sh., Ostendorf M., Richards C.* et al (2001), Normalization of Non-Standard Words, Computer Speech & Language, Vol. 15, pp. 287–333
8. *Sproat R.* (2010), Lightly supervised learning of text normalization: Russian number names, Workshop on Language Spoken Technology (SLT), IEEE, pp. 436–441
9. *Sutton C., McCallum A.* (2006), An Introduction to Conditional Random Fields for Relational Learning, MIT Press.

"VCHERA NASOCHINYALSYA VOROH STROK": PRODUCTIVE CIRCUMFIXAL INTENSIFYING PATTERNS IN RUSSIAN

Nedoluzhko A. Yu. (nedoluzko@ufal.mff.cuni.cz)
Charles University in Prague, Prague, Czech Republic

Khoroshkina A. S. (annakhor@gmail.com)
ABBYY; Lomonosov Moscow State University, Moscow, Russia

The current paper addresses verbal circumfixal derivation patterns in modern Russian. The discussion is focused on a series of circumfixes which trigger the intensified usage of the basic verb (~'*keep doing P too much*'). Derivatives built up by adding a prefix and a reflexive *-ся* to an imperfective verb are examined. Although each prefix adds specific shades of meaning to the verb, such patterns are, however, claimed to share common features at different levels of linguistic analysis, such as morphology, syntax, and semantics. Furthermore, such patterns are highly productive in modern language; once certain constraints are fulfilled, an intensified derivative can be formed from any imperfective verb. This fact, along with the patterns in question sharing certain common features, allow us to argue that they can be considered inflectional, rather than derivational.

Keywords: verbal prefixation, intensifying patterns, actional classes, semantics

"Майор полиции на Камчатке заигрался в Джеймса Бонда"
"Белорусский зайчик допрыгался до девальвации"[1]

1. Introduction

The productivity of circumfix derivation is a common linguistic phenomenon which is typical for many languages. Among other derivation models, patterns that derive new meanings using both a prefix and a suffix are common for Slavic languages, cf. Rus. *город* (*city*) → *пригородный* (*suburban*), *боль* (*pain*) → *обезболить* (*anaesthetize*), *смех* (*laugh*) → *насмехаться* (*mock*) and so on. In Hlaváčová-Nedoluzhko (2013), the series of Czech circumfixes with the prefixes *roz-, po-, za-, na-, vy-* and *u-* and the reflexive morpheme *se* together with their Russian equivalents were brought up in discussion. The question of whether a combination of a verbal prefix with the reflexive *-ся* can

[1] "A police major in Kamchatka has immersed into playing James Bond"; "The Belarusian zaichik ('bunny') jumped so hard that it ended up jumping into devaluation"

be called a circumfix is somewhat controversial; we will use this term below, as our point is that the prefix and the suffix are used to build up the pattern simultaneously.

Being added to a verb together with a reflexive morpheme, each of these prefixes forms a new meaning that specifies the semantics of the basic verb. However, this modification does not change the meaning of the verb itself, but rather its intensity. The productivity of this intensification pattern has been made use of for automatic lemmatization of verbs. However, its grammatical properties were not elaborated on in detail.

Upon observing these intensification patterns in Czech and comparing them to Russian, we can see that they are used similarly, but not in the same way. First of all, the prefixation derivation with *по-* does not require a reflexive morpheme in Russian. Moreover, the delimitative *по-* can be combined with other prefixes under analysis (cf. *понасмотреться*[2], *поиспи́саться* and so on). For these reasons, we do not consider this prefix part of our intensification pattern. On the other hand, unlike Czech and Slovak, another two Russian prefixes, *из-* and *до-,* together with the reflexive *-ся* may obtain the intensification meaning.

In our paper, we focus on Russian data in more detail. We argue that the prefixes *раз-, за-, на-, вы-, у-, из-* and *до-* together with the reflexive morpheme *-ся (-сь)* can combine with a large majority of imperfective verbs, forming new verbs with intensified meanings. Cf. the intensification pattern applied to the verb *плавать* (*swim*): *расплава́ться — запла́ваться — напла́ваться — вы́плаваться — упла́ваться — испла́ваться — доплава́ться*.

If represented in a diagram, the common semantics of the intensifying patterns looks as shown in Figure 1[3]:

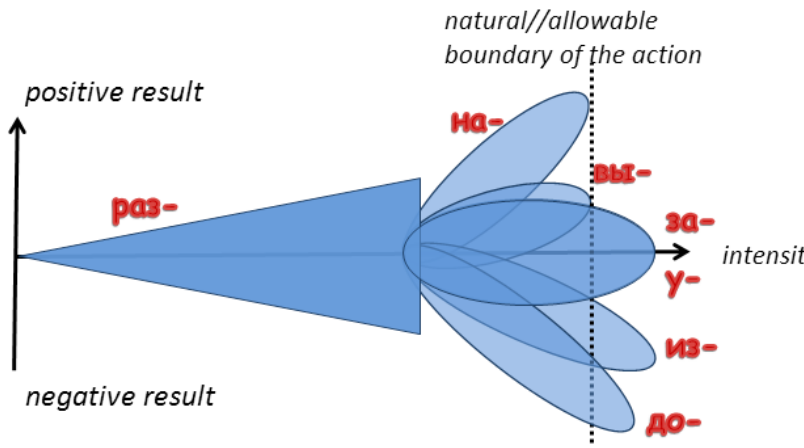


Fig. 1. Common semantics of the intensifying patterns

[2] This example is borrowed from Zaliznjak—Shmelev (2000). The authors claim that together with *понахвата́ться*, *понабра́ться* etc, this verb gets the intensified saturative meaning.

[3] This diagram is a modified version of the diagram presented in Hlaváčová—Nedoluzhko (2013).

According to the figure, the situation exceeds its natural intensity level and thus becomes abnormal for the patterns with the prefixes *за-, у-, из-* and *до-*. For instance, *заплаваться* literally means 'keep swimming for so long (or with such intensity) that swimming is already considered (either by the swimmer or by the speaker) as too intensive'. For the patterns with *на-* and *вы-*, the situation reaches the natural boundary of the action but does not exceed it. For the *раз-ся* circumfix, the meaning deals with the increasing intensity with no relation to the allowable boundary. We argue that the meanings of the circumfixal patterns are common for all the verbs to which they can be applied. We will also prove that all the circumfixes have common morphological, syntactic, and semantic features.

The paper is structured as follows: the shades of meaning for each prefix are specified in (3). In (2) and (4), respectively, we observe previous research on the subject and analyse the morphological features of the intensification pattern. The syntactic restrictions on the use of the intensification pattern are listed in (5). The main focus of our paper is the analysis of semantic constraints, see (6). The results and the conclusion are provided in (7).

2. Previous Research and the Aims of the Current Paper

Apart from the paper by Hlaváčová-Nedoluzhko (2013), whose ideas we are planning to extend in this work, our subject is addressed to in Zaliznjak—Shmelev (2000). The circumfix patterns in question are observed in their book within the word-formative category of action modes. The meanings of the patterns are classified according to different labelled modes and provided with examples, e.g. *раз- + -ся* derivation refers to ingressive and evolutive modes, *на- + -ся* refers to saturative mode, *до- + -ся* and *за- + -ся* refer to intensive-resultative mode. The circumfixes *вы- + -ся*, *из- + -ся* and *у- + -ся* are also mentioned within the intensive-resultative mode (Zaliznjak—Shmelev 2000, s. 106–118). Some useful remarks on meanings of the verbal derivatives in focus may be also found in Isachenko (1960).

Unlike academic dictionaries of Czech and Slovak[4], the Dictionary of Russian Language (1999) contains many common and not very common intensification meanings. For example, the dictionary describes the intensified usage of such relatively uncommon verbs as *набедствоваться, натолковаться, наборотьcя, набродиться, наваляться, нажиться*, etc. However, it does not include *набрызгаться, навертеться*, *наплеваться*. The saturative meaning of *наплаваться* is presented, but not *наныряться* and so on.

The most significant novelty of our work as compared to previous research is that we put these highly productive circumfixes together and examine common formal syntactic and semantic properties of the derivation pattern as a whole, rather than try to give an exhaustive semantic definition of each circumfix. We argue that it is the productivity of this pattern that explains the existence of a large number of derivatives not represented in dictionaries.

---

[4] As referred in Hlaváčová-Nedoluzhko (2013)

Our work addresses the field of verbal semantics, aspect and verbal actionality (aktionsart). For this reason, the classifications provided in Paducheva (1996, 2004) and Tatevosov (2002) are made use of (see 6.1.)

3. Definitions of Prefixes Combining with the Intensification Pattern

The following section presents an overview of the meanings[5] of the circumfixes under discussion. The letter P stands for an imperfective verb. By adding the given prefix and the reflexive morpheme to the verb we get the respective intensified form. Our definitions of the meanings are tentative. A more precise understanding can be obtained from examples that accompany every prefix entry.

раз-P-ся (with orthographic variants *разо-*, *разъ-* and *рас-*)

the action P, once started, has gradually increased and reached a high level of intensity, e.g. *разбаловаться, разлакомиться, разыграться*

…конь мой вороной ***разрезвился, расплясался, разыгрался*** *подо мной // my black steed* ***began romping about, dancing around, playfully jumping*** *under me*

за-P-ся

the action P has exceeded its natural or allowable boundaries, e.g. *заговориться, замечтаться, засидеться, зачитаться*

Засиделся *в гостях у печали, Что-то горек её крепкий чай // I* ***stayed way too long*** *at sadness' place, Oh how bitter is her strong tea*

на-P-ся

the action P has held for so long/with such intensity that its subject feels satisfied/annoyed with it, e.g.: *наговориться, нагуляться, насидеться, наслушаться*

Караул!! Ребенок ***наслушался*** *страшных историй!! // Oh no! The child* ***heard too many*** *scary stories!*

вы-P-ся

the action P has been carried out with such intensity that it has led to complete exhaustion of its (semantic) object, e.g.: *выплакаться, выспаться, выговориться*

Дождь под вечер ***выплакался*** *наспех, Скользкий ствол орешины намок // The rain* ***cried its eyes out*** *in the evening, the slick hazel stem got wet*

у-P-ся

the action P has been carried out with such intensity that it has led to an excessive result, e.g.: *упариться, упечься, упиться*;

---

[5] The meanings of the circumfixes are initially borrowed from the entries of the prefixes in the Dictionary of Russian Language (MAC, 1999) and then adapted to our corpus examples. The meanings in MAC for some prefixes (*за-*, *у-*) are also provided for non-reflexive verbs.

Норвежский турист ***упился*** *до состояния багажа // The Norwegian tourist* ***drank himself into*** *a stupor*

из-Р-ся

the action P has been carried out with such intensity that it has led to complete exhaustion of its subject (cf. *вы-Р-ся*), or to loss of a quality, e.g.: *извериться, изнервничаться, изолгаться*.

«Газпром» ***изнервничался****, ожидая, когда Украина назовет цену за «трубу» // Gazprom* ***were at their wit›s end****, as they waited for Ukraine to name a price for the "pipe"*

до-Р-ся

the action P has held for so long/with such intensity that it has led to a certain result, often a negative one, e.g.: *добудиться, дозвониться, дозваться*, *допрыгаться, доболтаться, добегаться*.

Первоклашка ***добегался*** *до перелома ноги // The first-grader* ***ran around so much*** *that he* ***ended up*** *with a broken leg*

The provided examples illustrate the intensified usage only. However, with the means of the verb intensification, verbs can be formed that already exist in common vocabulary of a language, but have a different meaning. Let's take the Russian verb *догадаться* as an example. In the Dictionary of Russian Language (1999) this lexeme is interpreted as 'make a guess, figure out by guessing'. This meaning is essentially different from the "intensified" meaning 'tell fortunes too much' which is formed by means of prefixation-postfixation and is not represented in the dictionary. Compare the following examples:

А тут девушка сама гадает. И ***догадалась*** *до того, что к ней пришел сам черт!* (internet) *// Here the girl was* ***telling fortunes*** *herself. And she* ***ended up*** *with the devil itself coming to her!*

Из речей девушек я ***догадался****, что дело шло о сыне соседки моей, богатой московской барыни. [П. Ю. Львов. Даша, деревенская девушка] // From the girls' talks I* ***guessed*** *that it was about my neighbor's, a rich moskovite lady's son.*

Other possible homonymic pairs are e.g. *разрешаться* ('resolve' and 'keep deciding too much'), *нажиться* ('make a fortune' and 'get tired of living somewhere for too long'), *извиниться* ('apologize' and 'be completely exhausted by blaming oneself') and so on.

4. Morphological Features of the Intensification Pattern

In Russian, there exist several ways of building derivatives with both a prefix and *-ся*, namely: adding one of the intensifying prefixes and the reflexive morpheme *-ся* to an imperfective non-reflexive verb (*решать—разрешаться)*; building

a decausative derivative from a perfective non-reflexive verb (*разрешать—разрешаться*); building a prefixal derivative from an imperfective reflexive verb (*решаться—разрешаться*). Below, we only appeal to derivatives that are built up by adding a prefix and a reflexive morpheme to an imperfective verb, and only in intensified usage. Because the original verb must be imperfective, the following verbs cannot be used as intensified: *извернуться* (*wriggle*), *наброситься* (*attack*), *раздаться* (*resound, expand*), *дотронуться* (*touch*), etc.

The resulting intensified verb is always perfective. For example, the verb *находиться* can have a saturative meaning only in perfective interpretation ('to walk too much'). If imperfective, it means 'be situated' or 'be found'.

It is crucially important that a verb to which the intensification pattern is applied already exists in the language. For this reason, intensification meanings are not available for such verbs as *изловчиться* (no or very rare *ловчить* or *ловчиться* in Russian), *измениться* (no **менить* or **мениться*), *наполниться* (no **полнить* or **полниться* in this meaning), *развернуться* (no **верзнуть* or **верзнуться*), *разлучиться* (no **лучить* or **лучиться* in this meaning) and so on. However, some exceptions can be occasionally found in the corpus[6]. Mainly, these are lexicalized usages that only confirm the rule. E.g. *До старта мы прожили в Гааге четыре дня и успели прилично* ***измочалиться****, загоняя себя на тренировках. [Наталья Бестемьянова и др. Пара, в которой трое] 'Before the start, we spent four days in Hague and* ***were completely exhausted with training****.'*

In case a circumfix is added to a verb that is already reflexive, *-ся* is not further doubled*: злиться* (be angry)- *разозлиться* (get angry), *смеяться* (laugh)—*насмеяться* (laugh enough), *купаться* (swim)—*укупаться* (be exhausted by swimming), *докупаться* (swim too much with a negative result) etc.

Intensifying derivation with a certain prefix can be applied to a verb which already contains this prefix (*заниматься* ('be occupied with smth')—*зазаниматься* ('be occupied with smth for a long time, with a possible negative result'), *разливаться* ('overflow')—*разразливаться* ('start overflowing with high intensity'), *заправляться* ('supply oneself with smth')- *зазаправляться* ('supply oneself with smth for a long time, with a possible negative result'), etc.).

Circumfixal derivatives with intensified meaning sometimes allow secondary imperfectivization (*разыграться* 'start playing (intensively), pfv'—*разыгрываться* 'start playing, imfv', *засидеться* 'sit for too long, pfv'—*засиживаться* 'sit for too long, pfv'). Supposedly, secondary imperfectivization of intensified derivatives is only possible under the condition that the same stem without the reflexive *-ся* and either without prefix or with another prefix (usually delimitative *по-*) allows such imperfectivization. E.g.: *засиживаться* is possible, as *сиживать* exists in the language, as well as *накуриваться: покуривать, дозваниваться: позванивать*, but not **размычиваться*, as **мычивать, *помычивать* (derived from *мычать*: 'moo').

---

[6] Russian National Corpus, http://ruscorpora.ru/en/index.html

5. Syntactic Constraints on the Intensification Pattern

5.1. Active Voice

The intensification pattern can usually only be applied to verbs in active voice. Thus, the intensification meaning is hardly available in case the reflexive morpheme is used as a passive or medial marker, e.g.:

Разговор ***раздробился****, запутался, и вскоре никому уже не было понятно, как вытаскивать загубленное предприятие. [Александр Солженицын. В круге первом, т. 1, гл. 26–51 (1968) // «Новый Мир», 1990]. // The conversation* ***got scattered****, mixed up, and it became already unclear how to save the ruined venture.*

Counterexamples are occasional and only possible in specific contexts. See, for example, the title of the current paper.

5.2. Object Generalization

Once the intensifying pattern is applied to a verb, it becomes intransitive. This is a purely syntactic constraint, as reflexive verbs in Russian are mainly intransitive. Cf. *читать книгу* (read the book)—*учитаться до смерти* (be totally exhausted by reading), but not **учитаться книгу до смерти* (be totally exhausted by reading the book). The semantic object can usually be expressed as an oblique object, in Genitive and sometimes also Instrumental case: *бросил камень — разбросался камнями; читал книгу — начитался книг*.

There is a remark in Zaliznyak and Shmelev concerning *на-* prefix: they point out that such derivatives can either take a Partitive object, or a combination with a quantitative word, such as 'many'. The latter possibility is not available for the intensifying pattern, but the former one is, as was shown above (*наловить рыбы\ наловиться рыбы*).

However, even if expressed within the same clause, the object usually gets generalized and loses its discrete semantics. One can say *разбросался камнями*, but never **разбросался камнем*, as one stone is something specific and cannot be used in the intensified context. For this reason, for instance, **убиться* is impossible in intensified usage: *бить* 'to beat' only allows a specific object.

Derivatives with *вы-* do not usually take oblique objects, given to the fact that the meaning of this circumfix (see 3) is that of complete exhaustion of its object. Neither do *из-* derivatives, for the reason that they imply complete exhaustion of the subject, thus the object is beyond the scope of their meaning.

Overall, the generalization of intensified derivatives' objects can be explained with the fact that the main focus of their meaning is the intensification of the action in current of time, which also implies iterativity and makes referring to a single or specific object troublesome.

6. Semantic Constraints on the Intensification Pattern

The set of semantic constraints for the intensifying pattern under discussion is somewhat more complicated.

On the one hand, we assume that the pattern is common for all the abovementioned prefixes, which is supported by the fact that all of them are subjects of the same morphological and syntactic constraints (see 4 and 5). Their common behavior seems to be natural if we agree that they all have similar semantics, namely that of intensification. In this chapter, we will try and prove that they do have certain semantic similarities as well.

On the other hand, it is beyond discussion that all prefixes have their specific meanings that overlap with the common intensification meaning and impose certain specific restrictions on the common pattern: for instance, one can say *налюбоваться*, but not **разлюбоваться*, as the verb *любоваться* ('admire the sight of') does not include the meaning of higher or lower intensity in its semantics. The prefix *раз-* requires such meaning from the verb, whereas *на-* does not.

We will provide an analysis of common distributive features of intensified derivatives, of the compatibility of the pattern with verbs of different actional classes, as well as of some semantic features not related to the class of a verb.

6.1. Compatibility with Verb Classes

One striking semantic feature of the intensifying pattern is its capability to combine with verbs of different classes with different results. Some classes appear to be highly productive, whereas the pattern can only occasionally be applied to the others. This is understandable given that the intensity of an action has to do with some limit that the action exceeds and thus it is crucially important what kind of lexical aspect class the verb belongs to.

In this section, we will examine this feature in more detail.

6.1.1. Telic Verbs

The intensifying pattern is only occasionally applied to all kinds of telic verbs. In such cases, a telic verb obtains the iterative meaning:

Половина зала — мужская, умирала от смеха, корчась. Один так ***разумирался****, что забыл, уходя, свой бумажник в отверстии для стаканов в кресле. [internet] // A half of the audience—the masculine one—was dying of laughter, making faces. One of them* ***got so much into dying*** *that he left his wallet in the glass-holder of his armchair upon leaving.*

Difficulties with deriving an intensified form from a telic verb are not without reason: the intensifying model, whatever the prefix is, always has a tint of exceeding some natural limit of the action. With telic verbs, a natural endpoint is included in the semantics of the verb and cannot be exceeded. Cf. in the example above the "natural

endpoint" of the verb *умирать* ('die') is the death itself, so one cannot "die with too much intensity, exceeding the limit". Therefore, such usages of telic verbs are occasional and are only felicitous in a specific context, like in the example above. Once applied to the verb, the intensifying pattern imparts an iterative meaning to it, turning one single action into a series of actions.

Thus, if an action cannot be iterated, it cannot undergo intensification: cf. **разлишаться (лишаться*: be deprived of).

6.1.2. Atelic Verbs

As atelic verbs do not include the semantics of a natural endpoint of the situation, they are more likely to be compatible with the intensification pattern. However, different classes of atelic verbs behave differently in this regard.

We have borrowed the classification of atelic verbs from Paducheva (1996). Her classification includes atemporal properties, inherent states, temporal states, processes, activities, occupations, and behaviors. Verbs are divided into classes according to a number of parameters. We have examined the ability of each class of atelic verbs to be intensified and obtained the following results:

- atemporal properties are not intensified[7], due to the fact that an atemporal property, according to its definition, cannot undergo any changes in process of time and thus cannot be a subject of the intensifying pattern.
- inherent states should be further divided into two classes: emotional states and others. While emotional states do get intensified (cf. *разгордился, заревновался)*, others do not. Another reason for separating emotional states from other states is that they usually combine with the delimitative prefix *по- (*which is one of the parameters of the classification).
- temporal states get intensified (in case all other conditions, such as presence of an animate subject, are successful, see below). Cf. *развеселился, изнервничался, намерзся*.
- processes do not get intensified for the reason described below: in Paducheva's classification, they are only processes with an inanimate subject, which contradicts the important constraint on the animacy of the subject.
- activities and occupations, unlike processes, tend to have an animate subject and are compatible with the intensification pattern. Cf. *расплакался, заработался, накувыркался, завоевался, испьянствовался, укомандовался*.
- behaviors do not usually undergo intensification except for *до-* and *на-*. The most probable reason for that is that behaviors, like atemporal properties, remain unchanged in current of time and thus do not include the meaning of higher or lower intensity in their semantics.

---

[7] However, due to the productivity of intensifying patterns (as was mentioned above), one can find examples of intensified derivatives of almost any verb, including atemporal properties, inherent states, and others, especially on internet blogs, cf.: Собака находится *весь день во дворе и это называется «гуляет»? Вы тот двор не видели…Угуляться можно — Неееее… Это не «угуляться»…. Это — «унаходиться» — Лишь бы не «удиваниться» в ожидании вечерней прогулки..*

6.2. Animate Subject Constraint

One interesting semantic feature of the intensifying pattern is that it only applies to verbs with an animate subject. The form *разболеться (болеть*—be sick) is quite common, whereas the form **размутиться (мутить*—feel sick, dizzy) is almost impossible because *мутить* is an impersonal verb.

At first glance, one can probably assume that this constraint has to do with the ability of the subject to control the action, i.e. with presence of an animate agent, but a more detailed analysis shows that this is not the case. Indeed, the pattern applies to agents (*разговориться: говорить*—talk), to experiencers (*расчувствоваться: чувствовать*—feel), and even to objects (*распадаться: падать*—fall down). What is important is that the subject, whatever theta-role it gets, is animate. Inanimate subjects do not allow intensifying: it is hard to imagine a form like ?*размерцаться (мерцать*—blink, twinkle) because a human being can hardly twinkle. Cf. also **Одежда насушилась (The clothes dried enough*: the example is borrowed from Tatevosov (2009)). When applied to a verb with an inanimate subject, the pattern slightly changes its meaning, i.e., the subject obtains traits of an animate creature, cf.: *вьюга разбушевалась (*the blizzard enraged *), лампа раскоптилась* (the lamp started smoking too much).

A confusing example of a dichotomy between a controlled and a noncontrolled action are pairs like *видеть — смотреть* (see—watch), *слышать — слушать* (hear—listen), where *засмотреться, заслушаться* are felicitous while **завидеться, *заслышаться* are not. Again, the first supposition that comes to mind upon looking at such pairs is that it depends on whether the action is controlled. However, as we have proved above, this parameter does not hold for all verbs, so another explanation must exist for this phenomenon. We suggest that the clue for it lies in the sphere of actional classes: *смотреть, слушать* are processes while *видеть, слышать* are states. We would be grateful to our readers for further suggestions on the subject.

6.3. Verbs of Oriented Motion

Verbs of oriented motion cannot be intensified: one can say *разлетаться, расходиться,* but never *разлететься, разойтись* in the intensified usage. The difference between *летать* and *лететь,* as well as between *ходить* and *идти* is that of oriented/non-oriented motion. *Летать* means to fly to and fro, while *лететь* means to fly in a specific direction. A similar observation has been made for one particular case, namely *до-ся,* in Zaliznyak&Shmelev (2000): they claim that this prefix is rarely combined with verbs that do not include the aim of the action in their semantics. We argue that this is true for all the prefixes in the scope of our discussion.

7. Conclusion and Perspectives

With our work, we wanted to highlight the fact that the circumfixes in question have a lot of common features at all levels of linguistic analysis, namely morphological and syntactical constraints, compatibility with different actional classes, the constraint on the animate subject etc.

Another striking peculiarity of the intensifying pattern is its productivity: even though we have tried and figured out certain semantic constraints on the original verb, while collecting data we kept facing the fact that intensified derivatives of almost any imperfective verb can be found on the internet, though those that violate the constraints may sound too colloquial.

These two features of the intensifying pattern allow us to argue that the category of intensification can be considered inflectional rather than derivational in modern Russian. The category fulfills the essential conditions on an inflectional model: it is productive and can be considered homogeneous with regard to its morphology, syntax, and semantics.

As we have pointed out above, the circumfixes differ slightly in their semantics: *раз-* triggers the meaning of gradual increase of intensity whereas *на-* and *у-* do not; *вы-* and *из-* differ in whether the object or the subject lies in the scope of the verb's meaning, and so on. We are planning to further develop the analysis of specific meanings of the circumfixes under discussion in the future.

One of the most exciting directions of our future research is comparing our results to other Slavic languages such as Czech, Slovak and Croatian. Some work in this direction has been already done, our task now is to carry out a more detailed semantic analysis of Czech and Slovak intensification patterns, which seem to be comparable to those in Russian but still have some challenging distinctions.

References

1. *Isachenko A. V.* (1960), Grammar of Russian language in comparison with the Slovak. [Grammaticheskii stroi russkogo yazyka v sopostavlenii s slovackim] V. II, Bratislava.
2. *Hlaváčová J.* (2009) Formal systems of Czech morphology with respect to natural language processing of Czech texts [Formalizace systému české morfologie s ohledem na automatické zpracování českých textů.]. Ph.D. thesis
3. *Hlaváčová J.* (2009) Verb intensification [Stupňování sloves] // After Half a Century of Slavonic Natural Language Processing. Masaryk University, Brno
4. *Hlaváčová J., Nedoluzhko A.* (2013), Intensifying Verb Prefix Patterns in Czech and Russian // Lecture Notes in Computer Science, Vol. 8082, Text, Speech and Dialogue: 16th International Conference, TSD 2013. Proceedings, Berlin / Heidelberg, pp. 303–310.
5. The Grammar of Russian Language (1980), [Grammatika Russkogo Yazyka]. Moscow: Izdatelstvo akademii nauk SSSR.

6. *Paducheva, E. V.* (1996), Semantic investigations [Semanticheskie issledovaniya]. Moskva, Yazyki russkoi kultury.
7. *Paducheva, E. V.* (2004), Dynamic models in lexical semantics [Dinamicheskie modeli v semantike leksiki]. Moscow: Yazyki slavianskoi kultury.
8. *Plungyan V. A.* (2011), Introducing grammatical semantics: Grammatical values and grammatical systems in the world's languages. [Vvedenie v grammaticheskuyu semantiku: grammaticheskie znacheniya i grammaticheskie sistemy yazykov mira] Textbook. Moscow, RGGU.
9. The Small Academic Dictionary of Russian Language (1999), [Malyj Akademicheskii Slovar russkogo yazyka.] Moscow: RAS Linguistic Studies Institute, Poligrafresursy.
10. *Tatevosov, S. G.* (2002), The parameter of actionality. Linguistic Typology 6(3): 317–401.
11. *Tatevosov, S. G.* (2009), Building Intensive Resultatives // Browne W. (ed.) Formal Approaches to Slavic Linguistics. The Cornell Meeting 2009. Ann Arbor: Michigan Slavic Publications, 289–302.
12. *Zaliznyak, Anna A., Shmelev, A.* (2000), Vvedenie v russkuyu aspektologiyu. Moskva.

A SUMMARIZATION MODEL BASED ON THE COMBINATION OF EXTRACTION AND ABSTRACTION

Osminin P. G. (osperevod@gmail.com)

South Ural State University, Chelyabinsk, Russia

We suggest a model of automatic summarization for scientific and technical texts. This model combines extractive and abstractive approaches for summarization and was developed on the basis of comparative analysis of authors' summaries and full texts of corresponding papers. The model consists of three main components: a keyword extractor, a domain and task oriented static knowledge base and a summarization algorithm. The keyword extractor is off-the shelf tool LanAKey_Ru, adapted to the application. Static knowledge includes stop lexicons, conceptual net, templates for summary content selection and rules for the generation. Stop lexicons are used for removing text segments irrelevant for the document summary. The conceptual net is used for semantic analysis of a document text helping content selection. Templates for information extraction are frame structures. Their slots are to be filled with extracted fragments of document sentences. Rules for summary generation define the grammar of summary sentences and their order. The summarization algorithm consists of four top level procedures—preprocessing, analysis, content selection and summary text generation. The model is described on the example of Russian scientific papers in mathematical modeling domain.

Keywords: automatic summarization, information extraction, knowledge base, conceptual net

1. Introduction

Automatic summarization allows quickly processing of great volumes of the scientific and technical documentation that is especially important in modern conditions when society encountered with the information overload problem.

Research in the field of automatic text summarization continue more than 55 years, but the problem of automatic creation of high-quality summaries still is not solved [29]. Automatic summarization methods can be divided into 3 groups: extractive methods, abstractive methods and hybrid methods.

Extractive methods [1, 9, 17, 21, 23, 26] create the text of the summary by extraction of the most significant text fragments (sentences, paragraphs etc.) of the source document. Fragments, as a rule, are extracted without changes and are inserted in the summary in the same sequence as in the source. The relevance of fragments can be defined by various criteria, for example by containing keywords, by fragment's location in the source text (titles, subtitles etc.), by presence of cue phrases [20], for example,

“the following results are obtained”, "it is important to note" etc. The advantages of extractive methods are relative independence of subject domain, lack of necessity of the detailed linguistic analysis and creation of extensive knowledge bases. Disadvantages of these methods consist in possible incoherence of summaries because text fragments of the source document are usually extracted without any processing. Also no information generalization is performed because words are not substituted by the more general concepts [11].

Abstractive methods [12, 19] create the text of the summary, as a rule, in three stages [30]: the source text is analyzed by means of the deep linguistic analysis, then an internal formal representation of the source text meaning is created, for example, in the form of frames, a semantic net, scripts etc. [7]. At this stage ontologies and knowledge bases of subject domains can be used. The compression of the source text meaning defines the summary content. The last stage is the generation of the summary text in a natural language [24].

Advantages of abstractive methods consist in providing more quality summary, than by extractive methods. The disadvantage is complexity for practical realization, abstractive methods need considerable amount of linguistic knowledge.

Due to complexity of abstractive methods the majority of automatic summarization methods are based on extraction of text fragments, and these methods show good results for specific types of texts [8].

Hybrid methods of automatic summarization are developed to overcome disadvantages of abstractive and extractive methods. In hybrid methods, the sentences (or their parts) are extracted from the source text and processed in different ways. For example, some parts of the sentences are omitted, some sentences are merged, sentences are inserted in the abstract in the other order as in the source etc. [25, 22]. Difficulties of hybrid methods developing consists in a choice of the most suitable combination of abstraction and extraction. In comparison with purely abstractive methods hybrid methods are easier to develop, in comparison with purely extractive methods hybrid methods can provide better quality of the resulting summary.

In automatic text summarization evaluation of the results is a very challenging problem. There are many works devoted to this problem [4, 14, 15, 16]. In [2] two groups of evaluation methods are distinguished—intrinsic and extrinsic. The intrinsic evaluation is oriented toward quality of the summary itself. For example, coherence of the text, its fluency, informativeness of the summary. The intrinsic evaluation often carrying out with participation of human experts who judge the quality of the summary comparing it with the gold standard (summaries created by human) or with results of other automatic summarization systems. However the problems of agreement between judges exist, as it is possible to create various summaries for the same text [6 18]. There are automatic evaluation methods, for example ROUGE [5].

The extrinsic evaluation assumes solving some task by means of the summary—for example, understanding of the full text by its summary. The judges are given a summary of the text and are asked some questions. Answers mean the comprehension of the full text. If the judge can provide answer the summary is considered correct.

In this paper we made attempt to contribute to a solution of one of important problems of automatic information processing and we present a hybrid summarization

model for scientific articles in Russian for “mathematical modelling” domain. The model is based on a combination of extractive and abstractive approaches. To our knowledge there are no hybrid automatic summarization methods for Russian. The model is created on the basis of the comparative analysis of full texts of scientific articles and corresponding authors' summaries.

2. The Comparative Analysis of Scientific Articles and Authors' Summaries

The analysis purpose consisted in detection of formal criteria of relevant fragments selection for the summary in the article’s text and choosing a technique used for summarization: extraction or abstraction, depending on what of these methods are used by the human in summary creation.

When analyzing overlaps of sentences of authors' summaries and sentences of corresponding articles the following parameters were considered:

- text presentation of the same content in the summary and article
- location of a relevant fragment for the summary in article
- lexical markers of the relevant information for the summary.

For analysis we select corpora of 107 full scientific articles (total amount of 203,729 word forms, without references) and corresponding authors' summaries (total amount of 4924 word forms), the average compression of full texts of articles was 41.4. After analysis of the material, we have found that the majority of authors’ summaries—54.3% is written by purely sentence extraction from the text. In 36.2% cases authors processed the extracted sentences from the text—paraphrased, omitted the unimportant information, added the new text, i.e. they combined extraction and abstraction. In 9.5% cases authors wrote summaries only of the new text (pure abstraction). As the majority of summaries are written by the authors from article’s sentences and/or the edited fragments of article we oriented our summarization model on a combination of extractive and abstractive approaches.

According to the requirements of the Russian state standard [13] in the summary text we can distinguish four informational parts: “Theme”—the information on a subject and an article’s theme, “Aim”—the information on the work purpose, “Method”—the information on a method or methodology of carrying out the work, “Result”—the information on results of work, a range of use of these results.

In the article’s text each of four mentioned types of the information, as a rule, is accompanied by lexical markers. To describe a theme of article markers “содержать” (to contain), “глава” (chapter), “раздел” (paragraph) etc. can be used. Markers “метод” (method), “инструмент” (tool), “находить” (to discover) etc. can be used to describe a method of research. We divided all lexical markers into the groups corresponding to specified informational parts—“Theme”, “Aim”, “Method” and “Result”. In each group markers are divided on the following semantic types: objects, attributes of objects, relations, attributes of relations. Objects describe a subject, relations describe relations between objects, attributes describe features of objects or relations.

Lexical, semantic-informational and morphosyntactic properties of markers are in a characteristic correlation. For example, the markers defining objects are expressed by nouns and pronouns; the markers defining the relations between objects are expressed by verbs and the markers defining attributes of objects and relations are expressed by adjectives, pronouns, adverbs. In the article's sentences markers can function as independent lexemes or can be a part of longer phrases. In the latter case markers according to a category are a part of noun phrases, verb phrases or adjective phrases containing terms of subject domain of mathematical modelling. For example, in a sentence "Рассмотрим следующую обратную задачу спектрального анализа", the marker "задача" (problem) is contained in a noun phrase "обратную задачу спектрального анализа".

3. A Model of Automatic Summarization

The model consists from three main components: a keyword extractor is off-the shelf tool LanAKey_Ru, the knowledge base and a summarization algorithm.

A keyword extractor LanAKey_Ru was developed by Sheremetyeva S. O. for extraction of keywords from patents in English, and then has been modified for extraction of keywords from articles on mathematical modelling in Russian [27, 28].

3.1. Knowledge Base

The knowledge base of automatic summarization model consists of the following main components: 1) stop-lexicons, 2) a conceptual net in the form of a rooted tree, 3) sets of templates for information extraction and 4) rules of the analysis of the full document, generation of the summary text, arrangement of sentences in the summary text.

The stop-lexicons are used for removing from the full text of article the irrelevant information for the summary for the purpose of facilitation of the further analysis. The stop-lexicons consist of three lists:

a) the list of the stopwords which are deleted (for example, "итак" (so), "однако" (however), "окончательно" (finally)),
b) the list of the words defining a part of a sentence for removing (for example if in a sentence the word "где" (where) is found then the sentence part after this word is deleted),
c) the list of the words defining a sentence for removing (for example, "положим" (let us assume), "пусть" (let), "обозначим" (let us define)).

The conceptual net is used for the semantic analysis—detection of lexical markers in the document text. The net consists of terminal and non-terminal nodes and the links realizing the relation of inclusion. The root of the tree is concept "Summary", in non-terminal nodes there are concepts, corresponding to informational parts of the summary "Theme (T)", "Aim (A)", "Method (M)", "Result (R)" and to semantic types of markers "Object (O)", "Relation (P)", "Attribute of object (AO)" and "Attribute of relation (AP)". Concepts reflect the required types of the information for summary.

Terminal nodes—lexical units—are the markers, realizing these concepts in the text. Lexicons of markers include all word forms of marker lexemes that have been selected in the course of the analysis of a sublanguage of mathematical modelling.

The fragment of conceptual net for a node “Method (M)” is showed on Fig. 1.

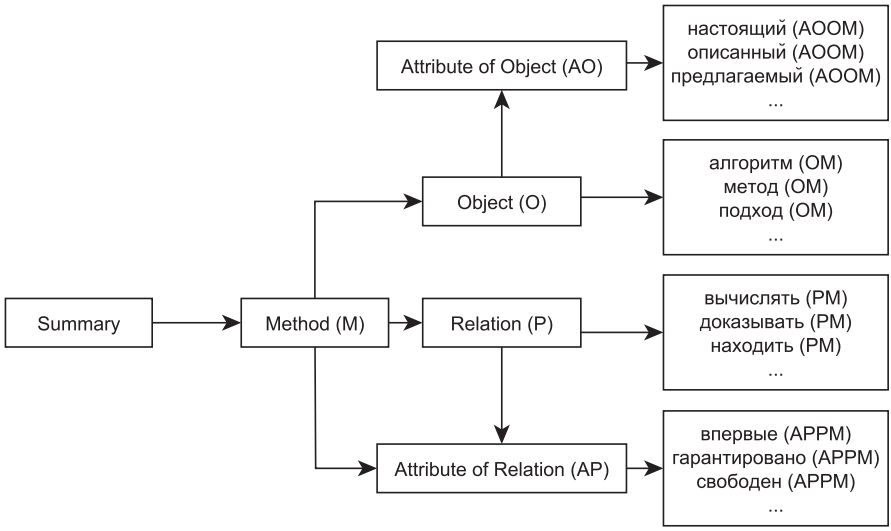


Fig. 1. A fragment “Method (M)” of a conceptual net

Templates for information extraction are divided into four categories—Theme, Aim, Method, Result, according to type of the extracted information required in accordance with the Russian state standard. Templates are frame structures of a following kind:

Template	::= (IP (structure))
IP	::= {theme, aim, method, result}
Structure	::= (X Phrase X … Phrase …X)
X	::= (word word ... word)
Phrase	::= {NP(MARKER T), VP(MARKER T), AP(MARKER T)}
Marker	::= (marker(net code))
Template number	::= (natural number)
Sentence number	::= (natural number)
Weight	::= (natural number)

where,

IP—informational part of the summary; structure—structure of an article’s text fragment; X—chain of consecutive words of a fragment, can be empty; marker—a terminal node of the net; code—the net code; T—the term (can be empty); NP—noun phrase, VP—verb phrase, AP—adjective phrase, template number—a sequence number

of a template which is defined at template filling, sentence number—a sequence number of the sentence used for filling a template, weight—template weight is calculated at a template weighing. If some part of a template is in square brackets it can be empty.

Rules of text generation represent the text fragments in filled slots of template in the coherent sentences. From each template, as a rule, one sentence is generated. If templates belong to the same sentence or begin with identical words ("рассматривается" (is considered), "в работе" (in the work)) then from these templates one sentence is generated. Grammatical rules include:

1) Verbs of the first person are substituted with impersonal forms (the third person with the ending "-ся"), the further noun is put in the nominative case.
2) Titles of article parts (paragraph, section, table etc.) can either be omitted, or can be substituted by expression "в статье" (in paper), "статья" (paper).
3) In templates Theme, Aim, Method the verb is used in the present. In a template Result the verb is used in the past tense (except verbs "является" (is), "смогут" (can), "позволят" (will allow)).

The order of generated sentences in the summary text is defined by the following rule:

1) Sentences follow in this order Theme, Aim, Method, Result. If in a category there are more than one sentence they are ordered by weight which is calculated from weight of the keywords and weight of the markers.

3.2. Automatic Summarization Algorithm

The algorithm consists of 4 main procedures—preprocessing of the article's text, the analysis of the article's text, content selection for a summary, generation of the summary text. The main procedures include subprocedures. Each procedure receives some information as an input and produces result which is used as the input for the following procedure.

The preprocessing procedure consists of three consecutive subprocedures—sentence segmentation, compression of the text, extraction of keywords.

The first subprocedure performs a sentence segmentation of the article's full text. We consider the sentences as text segments from a dot to a dot.

Then text compression subprocedure removes from the text irrelevant fragments for the summary. At this step the stop-lexicons from the knowledge base are used. After fragments' removing the sentences containing less than five words are removed (a word is a text fragment from a space to a space). Text compression is intended for facilitation of the further analysis.

After compression the extraction of keywords from the text by means of extractor LanAKey_Ru is performed. LanAKey_Ru is capable to extract nominal phrases up to four words without a preliminary annotation of the text. Keywords in our model are the most relevant noun phrases (NP) of the article. Relevance is calculated according to empirically found formula $(d / D) + U * 10$, where d—number of sentences where NP occurred at least once, D—number of sentences in the text, U—uniqueness,

shows that the keyword functions individually, instead being a part of a longer phrase. This parameter is calculated as a difference between frequency of NP and the sum of frequencies of longer noun phrases containing the given NP. All specified parameters are defined in the extractor LanAKey_Ru. From the text of a full article the 10 most relevant keywords are extracted.

Analysis procedure consists of two subprocedures—partial morphosyntactic analysis and the semantic analysis. At a stage of morphosyntactic analysis detection of lexical groups and their part of speech tagging is performed. The morphosyntactic analysis is performed as follows. The phrases coinciding with keywords are tagged with noun phrase tags and are assigned the weight (relevance) automatically defined by LanAKey_Ru. Thus, LanAKey_Ru not only extracts the keywords and define their relevance but also performs an essential part of the morphological analysis, since NP—the most frequency lexical group in any text type. Then morphosyntactic analysis is finished by means of the software of Aot.ru project [3].

In such annotated article's text the semantic analysis is carried out by means of conceptual net. The lexicon from the terminal nodes of a net is compared to the annotated text. When the coincidence occurs the marker is given the network code—a path from a terminal node to the net vertex. For example, the net code for a marker "подход" (approach) is OM (Object-Method), a net code for a marker "prove" (доказывать)—PM (Relation-Method). Due to homonymy of markers—the same word forms can express various relations—during semantic analysis one marker can be given various net codes. The resolution of marker ambiguity takes place by means of templates at later stage of the analysis.

The third procedure—content selection consists of two subprocedures—scoring of sentences' weight (relevance) and filling of templates. The sentence weight (relevance) is calculated by the following formula:

$$W = 10N + M_i + K_i$$

where

W—weight of the sentence (relevance),

N—number of keywords in a sentence, the multiplier 10 was found empirically,

M_i—weight of all markers in a sentence (the weight of one marker is 10),

K_i—weight of all keywords in a sentence (weight is taken from Lana-key_Ru)

We define that for small texts (up to 9000 characters with spaces) the selection threshold is 10 most relevant sentences, for the bigger texts—10% of the most relevant sentences.

Filling of templates subprocedure begins with review of the selected sentences from left to right. In templates for information extraction the order of markers and other words in a sentence is set. If the sentence (or its part) satisfies template requirements then template slots are filled with sentence parts. Applying of templates on text fragments occurs between punctuation marks.

For example, the following sentence from the article's text: *В этом параграфе изложены используемые при получении основного результата факты из классической теории полугрупп операторов* satisfies the template Theme

from the knowledge base *(Theme [X] [(AP(Marker(AOT) T)] (NP(Marker(OT) T)) [(AP(Marker(APT) T)] (VP(Marker(PT) T) X))*. The result of this template filling by sentence fragments is shown below:

IP	::= Theme
X	::= В
AP(Marker(AOT) T	::= этом
NP(Marker(OT) T	::= параграфе
AP(Marker(APT) T	::=
VP(Marker(PT) T	::= изложены
X	::= используемые при получении основного результаты факты из классической теории полугрупп операторов
Template number	::= 2
Sentence number	::= 7
Weight	::= 6

The order of template applying is the following Theme, then Aim, Method, Result.

The example of formally generated summary for article [10] and its author's summary is shown below.

Formally generated summary

В статье рассматривается задача Коши для абстрактного линейного эволюционного уравнения с памятью в банаховом пространстве. В статье изложены используемые при получении основного результата факты из классической теории полугрупп операторов.

С помощью принципа сжимающих отображений доказана однозначная локальная разрешимость этой задачи в смысле классических решений.

Результаты работы позволят с одной стороны перейти к рассмотрению полулинейных эволюционных уравнений с памятью. Полученный результат использован при исследовании начально-краевой задачи для параболического интегродифференциального уравнения с памятью. Основным результатом данной работы является доказательство однозначной локальной разрешимости этой задачи.

The author's summary

Доказана локальная однозначная разрешимость задачи Коши для линейного эволюционного уравнения с секториальным оператором и с интегральным оператором памяти в банаховом пространстве. Результат работы проиллюстрирован на примере начально-краевой задачи для интегро-дифференциального уравнения с частными производными.

Formally generated summary reflects a content presented in the author's summary and satisfies the requirements of the Russian state standard in a greater degree. The evaluation of the results was performed manually and consisted in comparison of formal summaries with authors' summaries in order to detect the informational

parts required in accordance with the Russian state standard. In case of considerable contradiction between the formal and author's summary we resorted to help of the experts. As a result we discover that about 70% of formal summaries was generated correct. Comparison with automatic summarization systems was not performed, as we did not have a possibility to get the data of such systems for comparison.

4. Conclusion

In this paper we presented the model of automatic summarization, developed on the basis of comparative analysis of authors' summaries and full texts of corresponding articles.

In the presented model extractive and abstractive approaches are realized. We described the model parts: 1) keyword extractor, 2) the knowledge base and 3) algorithm of automatic summarization. Steps of automatic summarization algorithm are developed. Formally generated summaries, as a rule, coincide with authors' summaries by content and satisfy the requirements of the Russian state standard in a greater degree.

References

1. *Abuobieda A., Salim N., Albaham A. T., Osman A. H., Kumar Y. J.* (2012), Text summarization features selection method using pseudo genetic-based model, International conference on information retrieval knowledge management, Kuala Lumpur, pp. 193–197.
2. *Afantenos Stergos, Karkaletsis Vangelis, Stamatopoulos Panagiotis* (2005), Summarization from medical documents: a survey, Artificial Intelligence in Medicine. Vol. 33 Issue 2, pp. 157–177.
3. Automatic text processing [Avtomaticheskaja Obrabotka Teksta], available at: http://aot.ru/
4. *Chin-Yew Lin, Eduard Hovy* (2002), Manual and Automatic Evaluation of Summaries, Association for Computational Linguistics. Proceedings of the Workshop on Automatic Summarization (including DUC 2002), Philadelphia, pp. 45–51.
5. *Chin-Yew Lin* (2004), ROUGE: A Package for Automatic Evaluation of Summaries, Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain, pp. 74–81.
6. *Das D., Martins Andre F. T.* (2007), A Survey on automatic text summarization. Unpublished manuscript, Literature survey for Language and Statistics II, Carnegie Mellon University, 31 p.
7. *DeJong G. F.* (1982), An Overview of the FRUMP System, in Strategies for Natural Language Processing, Lawrence Erlbaum, Hillsdale, New Jersey, pp. 149–176.
8. *Dubinina E. Ju.* (2013), Scientific text compression: methods and models [Kompressija nauchnogo teksta: metody i modeli], Avtoreferat dis. … kandidata filologicheskih nauk: 10.02.21, Rossijskij gosudarstvennyj pedagogicheskij universitet im. A. I. Gertsena.

9. *Edmundson H. P.* (1969), New methods in automatic extracting, Journal of the ACM, vol. 16 Issue 2, pp. 264–285.
10. *Fedorov V. E., Staheeva O. A.* (2008), On Local Solvability of Linear Evolutionary Equations with Memory [O lokal'noj razreshimosti linejnyh èvoljutsionnyh uravnenij s pamjat'ju], Bulletin of the South Ural State University. Series Mathematical Modelling, Programming & Computer Software [Vestnik Juzhno-Ural'skogo gosudarstvennogo universiteta. Serija: Matematicheskoe modelirovanie i programmirovanie] no. 27 (127), pp. 104–109
11. *Genest Pierre-Etienne, Lapalme Guy, Yousfi-Monod Mehdi* (2009), Hextac: the creation of a manual extractive run, Proceedings of the Second Text Analysis Conference, Gaithersburg.
12. *Genest Pierre-Etienne, Lapalme Guy* (2012), Fully Abstractive Approach to Guided Summarization, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, pp. 354–358.
13. GOST 7.9-95 (1995), System of standards on information, librarianship and publishing. Informative abstract and indicative abstract. General requirements [GOST 7.9-95. Sistema standartov po informatsii, bibliotechnomu i izdatel'skomu delu. Referat i annotatsija. Obshchie trebovanija] Izdatel'stvo standartov, Moscow.
14. *Hassel Martin* (2004), Evaluation of Automatic Text Summarization. A practical implementation Licentiate Thesis Stockholm, Sweden, Royal Institute of Technology (KTH) 69 p.
15. *Hirao Tsutomu, Okumura Manabu, Yasuda Norihito, Isozaki Hideki* (2007), Supervised automatic evaluation for summarization with voted regression model. Vol. 43 No 6, pp. 1521–1535.
16. *Hobson Stacy President, Dorr Bonnie J., Monz Christof, Schwartz Richard* (2007), Task-based Evaluation of Text Summarization Using Relevance Prediction, Information Processing and Management. Vol. 43 No 6, pp. 1482–1499.
17. *Jatsko V. A.* (2002), Symmetrical Summarization: Theoretical Foundations and Methods [Simmetrichnoe referirovanie: teoreticheskie osnovy i metodika], NTI. Ser. 2 no. 5. pp. 18-28.
18. *Karen Sparck Jones* (2007), Automatic summarising: The state of the art, Information Processing and Management, vol. 43 issue 6, pp. 1449-1481.
19. *Korhova O. V.* (2001), Method for mathematic formalization of the Russian language in automatic text summarization aspect [Metod matematicheskoj formalizatsii russkogo jazyka v zadache avtomaticheskogo referirovanija tekstov], dis. kand. fiz-mat. nauk: 01.01.09.
20. *Leonov V. P.* (1986), Summarization of scientific and technical literatre [Referirovanie i annotirovanie nauchno-tehnicheskoj literatury], Nauka, Novosibirsk.
21. *Luhn H. P.* (1958), The Automatic Creation of Literature Abstracts, IBM Journal of Research and Development, vol. 2, no. 2, pp. 159–165.
22. *Polanyi Livia, van den Berg Martin H., Lorenzo Thione Giovanni, Crouch Richard S., Culy Christopher D., Ahn David D.* (2009), Systems and methods for hybrid text summarization, United States Patent US 7,610,190 B2

23. *Prihod'ko S. M., Skorohod'ko È. F.* (1982), Automatic summarization based on inter phrase links [Avtomaticheskoe referirovanie na osnove analiza mezhfrazovyh svjazej], NTI. Ser. 2 no. 1. pp. 27–32.
24. *Radev Dragomir R., McKeown Kathleen R.* (1998) Generating Natural Language Summaries from Multiple On-Line, Computational Linguistics—Special issue on natural language generation, vol. 24, no. 3, pp. 470–500.
25. *Saggion Horacio, Lapalme Guy* (2002), Generating indicative–informative summaries with SumUM, Computational Linguistics. vol. 28, no. 4, pp. 497–526
26. *Sevbo I. P.* (1969), Structure of coherent text and automatization of summarization [Struktura sviaznogo teksta i avtomatizatciia referirovaniia], Nauka, Moscow.
27. *Sheremetyeva S.* (2009), An efficient patent keyword extractor as translation resource, MT Summit XII: Third Workshop on Patent Translation, Ottawa, pp. 25–32.
28. *Sheremetyeva S.* (2012), Automatic Extraction of Linguistic Resources in Multiple Languages, Proceedings of NLPCS 2012, 9th International Workshop on Natural Language Processing and Cognitive Science in conjunction with ICEIS 2012, Wroclaw, pp. 44–52.
29. *Sheremetyeva S. O.* (2013), On interactive summarization oriented to machine translation [Interaktivnoe referirovanie, orientirovannoe na mashinnyj perevod], Bulletin of the South Ural State University Series Linguistics [Vestnik Juzhno-Ural'skogo gosudarstvennogo universiteta. Serija: Lingvistika], vol. 10, no. 1, pp. 89–92.
30. *Tarasov S. D.* (2010), Modern methods for automatic summarization [Sovremennye metody avtomaticheskogo referirovaniia], St. Petersburg State Polytechnical University Journal. Computer Science. Telecommunications and Control Systems [Nauchno-tehnicheskie vedomosti Sankt-Peterburgskogo gosudarstvennogo politehnicheskogo universiteta. Informatika. Telekommunikatsii. Upravlenie] no. 6 (113) pp. 59–74.

СНЯТАЯ УТВЕРДИТЕЛЬНОСТЬ И НЕВЕРИДИКАТИВНОСТЬ[1]

Падучева Е. В. (elena.paducheva@yandex.ru)

ВИНИТИ РАН, Москва, Россия

Ключевые слова: снятая утвердительность невериди кативность отрицательная поляризация нереферентность

SUSPENDED ASSERTION AND NONVERIDICALITY

Paducheva E. V. (elena.paducheva@yandex.ru)

VINITI RAN, Moscow, Russia

Two notions are compared: *suspended assertion* and *nonveridicality*. It is argued that these notions, though used in the frameworks of different linguistic theories, are applied to similar linguistic phenomena. In this paper the notion of nonveridicality is applied to one group of Russian indefinite pronouns — namely, to **negative polarity pronouns** (NPP). Four groups of non-referential indefinite pronouns are differentiated in Russian: negative pronouns (*ni-* series), non-specific indefinite (*-nibud'* series), free choice (*ugodno* series and *ljuboj*) and negative polarity pronouns (*-libo* and *by to ni bylo* series). Following Giannakidou 1998, I reject the hypothesis that NPPs are licensed in the context of downward entailment operators only. I also argue against what is claimed in Giannakidou 2011, that NPPs are licensed in the three types of environments: negative, downward entailing and nonveridical: all contexts of the Russian NPPs can be demonstrated to be nonveridical, and the context of negation is one of them. The list of contexts licensing all the four classes of non-referential pronouns is suggested. Each of the four classes of pronouns chooses its own subset of contexts from the list.

Key words: suspended assertion nonveridicality negative polarity non-referentiality

[1] Данная работа была выполнена при финансовой поддержке РГНФ, грант № 14-04-00604а.

Термин *снятая утвердительность* (suspended assertion) принадлежит У. Вейнрейху (Weinreich 1963/1970: 173). Он использовался в Падучева 1985; языковые факты, которые требуют обращения к этому понятию, обсуждались далее в Падучева 2004, 2005, 2011, Богуславский 2001, 2008, Борщев и др. 2008, Подлесская 2012 и др.

С конца 90-х годов прошлого века в англоязычной лингвистической литературе получает распространение термин *nonveridicality* (предлагаемый перевод на русский язык — *невередикативность*), см. Zwarts 1995; Giannakidou 1998, 2002, 2006; Karttunen, Zeanan 2005 и мн. др. Эти работы рассматривают примерно тот же круг явлений, которые охватываются термином *снятая утвердительность*, но в рамках другой теории — в формальной семантике. Возникает задача сравнить результаты, полученные при разных подходах: речь идет о важном аспекте семантики — о референции пропозициональных компонентов предложения.

Явления, о которых идет речь, принадлежат к сфере грамматической модальности, где отсутствует общепринятая терминология; так что следует, прежде всего, зафиксировать терминологию, относящуюся к этой сфере.

1. Три семантических сферы в зоне грамматической модальности

В статье Модальность в http://rusgram.ru/index различаются следующие три типа грамматической модальности:

1) **объективная модальность**, т. е. статус ситуации по отношению к реальному (так сказать, объективному) миру; он выражается, в частности, противопоставлением, индикатива и сослагательного наклонения; эта модальность может быть **реальная** (как в *Никита ему уже позвонила*), **контрфактивная** (как в *Если бы Никита вернулась, она бы ему уже позвонила*) и **нейтральная** (как в *Думаю, Никита ему уже позвонила*);

2) **иллокутивная модальность**, т. е. коммуникативная цель, которую ставит перед собой говорящий в своем высказывании; различается **утвердительная** иллокутивная модальность, **побудительная** и **вопросительная**; есть четвертая возможность — пропозиция может быть лишена иллокутивной силы, т. е. иметь **нулевую иллокутивность**; такая возможность возникает только в синтаксически подчиненной позиции;

3) **субъективная модальность**, т. е. отношение говорящего к ситуации — эмоциональная, ментальная, волитивная или еще какая-то установка); она выражается, в русском языке, сослагательным наклонением в оптативном значении (как в *Хорошо бы наши выиграли*!), показателями эпистемической возможности и необходимости (*Он мог забыть о нашей договоренности*), вводными словами и другими средствами (*Едва ли он согласился*); субъективных модальностей очень много.

Эти три вида модальности тесно связаны друг с другом. Так, пропозиция, которая имеет не утвердительную, а побудительную или вопросительную иллокутивную модальность, имеет нейтральную реальную модальность, т. е. употребляется **безотносительно к истине**. То же может происходить с пропозицией в контексте субъективной модальности.

2. Снятая утвердительность по Вейнрейху

В Weinreich 1963/1970 индикатив называется утвердительным наклонением — в самом деле, индикатив выражает утвердительную иллокутивную модальность. Перечисляются средства, которые используются в разных языках для того, чтобы снять утвердительность, изначально присущую индикативу. Это показатели **снятой утвердительности** (assertion suspending devices), иначе — языковые средства нейтрализации утвердительности (neutralization of assertiveness, или suspension of assertion). К ним принадлежит императив и другие косвенные наклонения, такие как конъюнктив, который выражает «прямой отказ от ответственности за истинность высказывания» или эвиденциальность; буд. время; ту же роль выполняют номинализованные и инфинитивные конструкции. Эти показатели создают для пропозиции **контекст снятой утвердительности**. Такой же контекст создают модальные слова (типа *может, хочет, должен*, *необходимо*), отрицание, вопрос, дизъюнкция, целевые и условные союзы; предикаты пропозициональной установки, выражающие неуверенность, предположительность, нереальность.

*Термин Вейнрейха suspen*ded *assertion* был переведен на русский язык как «снятая утвердительность». Этот перевод соответствует уважаемой логической традиции. Так, А. Черч различает «утвердительное употребление предложений, с одной стороны, и, с другой стороны, **неутвердительное употребление** их в качестве <…> частей более длинных предложений» (Черч 1960: 30).

Черч широко пользуется термином *proposition* — в значении ‘смысл предложения’. В русском переводе 1960 года англ. *proposition* переводится как *суждение*, но сейчас, без сомнения, в переводе было бы использовано слово *пропозиция*, которое широко употребляется в современной лингвистике именно в нужном Черчу значении: англ. слово *proposition* Черч называет счастливым результатом процесса, в результате которого удалось избежать смешения предложений и их смыслов (Черч 1960: 32).

В Черч 1960: 357 говорится, что «неутвердительно употребляемые предложения всегда являются частями утверждаемых предложений». Это верно, если ограничиться речевым актом утверждения и не рассматривать других типов речевых актов — побуждения и вопроса, в составе которых предложение тоже употребляется безотносительно к истине, т. е. не является ни истинным, ни ложным.

О том, что пропозиция — это смысл предложения и что пропозиция обретает истинностное значение тогда, когда она утверждается или используется в каком-то другом контексте, идет речь в Vendler 1967, 710: “propositions *per se* are neither true nor false. Only those *held* in a certain way (e. g. belief, opinion), *issued* with a certain force (e,g, statement, verdict) or *viewed* in a certain context (e. g. as expressing a fact) are true or false.”

Лингвистическая семантика, если она ставит целью адекватное представление смысла предложений языка, должна оперировать не только с утверждаемыми пропозициями (которые являются истинными или ложными), но и с пропозициями, которые лишены параметра истинности. Это может быть, напр., пропозиция, которая имеет одну из двух неутвердительных иллокутивных модальностей, вопросительную (*Ушла ли Маша?*) или побудительную (*Уходи,*

Маша!); пропозиция в синтаксически подчиненной позиции, которая лишена самостоятельной иллокутивной силы, т. е. имеет нулевую иллокутивность, как у ‘Маша ушла’ в контексте *<Я думаю, что> Маша ушла*; ассоциированная пропозиция вводного слова: *Маша, возможно, ушла*.

По Вейнрейху, утвердительная иллокутивность (иначе — ассертивность) — необходимое условие для того, чтобы пропозиция с глаголом в индикативе была соотнесена с реальностью, т. е. обозначала ситуацию реального мира: пропозиция в контексте снятой утвердительности лишена параметра истинности, т. е. не имеет ни реальной, ни ирреальной объективной модальности. На самом деле, тут есть один недосмотр: в контексте фактивного или импликативного глагола /предикатива, когда пропозиция оказывается презумпцией или импликативом, она, будучи лишена иллокутивной силы, имеет, однако, объективную модальность истина. Например, в контексте *Мне жаль, что Маша ушла* пропозиция ‘Маша ушла’ истинна: она лишена иллокутивной силы, но не объективной модальности. Поскольку термин «снятая утвердительность» получил достаточное распространение в значении безотносительность к истине, я оставляю за собой право употреблять его в этом, уточненном, значении, хотя оно не вполне соответствует его внутренней форме.

В Падучева 1985: 33, 94f, 215–220 понятие снятой утвердительности было использовано при описании русских **нереферентных неопределенных местоимений** (non-specific indefinite, NSI), типа *какой-нибудь, кто-либо, какой бы то ни было, любой*. Они практически недопустимы в утвердительном контексте, ср. **Он купил что-нибудь* (равно как и в фактивном: **Хорошо, что он купил что-нибудь*), но употребляются в контексте снятой утвердительности — в косвенных наклонениях и в гипотаксической позиции: *Купи что-нибудь*, *Куплю что-нибудь*, *Он купит что-нибудь*? *Купить что-нибудь*, *Если он купил что-нибудь*, *Могу купить что-нибудь* и т. д. Они возможны в контексте вопроса, императива, буд. времени, сослагательного наклонения, условия, нереальной модальности, дизъюнкции, дистрибутивности, узуальности. В предложении *Кто-нибудь ей помог* скрытая субъективная модальность: ‘наверно, кто-нибудь’.

В докладе Падучева 2004 был представлен ряд других явлений, обусловленных контекстом снятой утвердительности: исчезновение семантических актантов у некоторых глаголов в прямой — не параметрической — диатезе; широкая сфера действия отрицания в глагольно-адвербиальном комплексе и в предложениях с кванторными словами и др. В данном докладе речь идет о снятой утвердительности в применении к местоимениям с отрицательной поляризацией. При этом я опираюсь на вышеупомянутую серию работ по формальной семантике, использующих понятие (не)веридикативность.

3. Неверидикативность и отрицательная поляризация

Определение (не)веридикативности рассматривает контекст как пропозициональный оператор. Пропозициональный оператор (или контекст) *F* является для пропозиции *p* **веридикативным**, если и только если *Fp* имеет следствием или

пресуппозицией *p*; в противном случае оператор (или контекст) *F* является **неверидикативным** (см. Zwarts 1995, Giannakidou 1998). Неверидикативность — это то же, что снятая утвердительность в уточненном варианте определения.

В статье Zwarts 1995 неверидикативность используется при анализе слов и оборотов с **отрицательной поляризацией** (negative polarity items, сокращенно — NPI), в частности, английского местоимения *any*. В центре внимания — ограничения сочетаемости, которые свойственны этим единицам, т. е. контексты их употребления.

Упоминание об NPI есть уже в классической статье об отрицании Klima 1964: NPI употребляются преимущественно в контексте отрицания, но также и в некоторых других контекстах — например, в вопросе или в условном предложении. В рамках формальной семантики были попытки выявить семантическую мотивацию для этих ограничений сочетаемости. В Ladusaw 1980 была выдвинута гипотеза, что все контексты, допускающие NPI, объединены следующим общим свойством: это контексты операторов «выводящих вниз» — downward entailing operators. Оператор называется **выводящим вниз**, сокращенно — **DE-оператором**, если в контексте этого оператора из истинности некоторого утверждения относительно некоторого множества следует истинность этого утверждения относительно подмножества этого множества. Например, *немногие* и *редко* — это DE-операторы:

(1) а. Немногие эскимосы едят овощи
б. Немногие эскимосы едят шпинат;
(2) а. Джон редко ест овощи.
б. Джон редко ест шпинат.

Со временем стало ясно, что DE-операторы не оправдывают возлагавшихся на них надежд. Например, DE-операторы не работают в контексте вопроса. Были попытки усиления или ослабления определений, но в конце концов было признано, что DE-операторы охватывают лишь часть контекстов, лицензирующих NPI *any*. В статье Zwarts 1995 и в многочисленных работах A. Giannakidou возникает идея о том, что условием, лицензирующим NPI, является также контекст неверидикативности.

То, что DE-операторы не охватывают всех контекстов употребления NPI, — это только часть дела. Главная претензия Джаннакиду к гипотезе Лэдюсоу — в том, что она претендует на то, чтобы быть семантической, между тем, 1) непонятна семантическая связь между DE-контекстами и семантикой NPI: почему DE-контекст должен быть притягателен для NPI; 2) непонятно, почему контексты могут быть разными для разных NPI в одном языке и для NPI с близкой семантикой в разных языках. Согласно Джаннакиду, убедительный ответ на эти вопросы дает неверидикативность.

1) Общим свойством отрицательно поляризованных единиц является их **референциальная неполноценность**: NPI — это именные группы с **ограниченными референциальными возможностями**: они не обозначают конкретных объектов и не вводят их в контекст речевого акта. Отсюда тяготение этих ИГ к контекстам неверидикативности.

2) Сочетаемость той или иной языковой единицы определяется ее семантикой. При наличии какой-то общности, она может быть разной у близких по смыслу NPI — как в одном языке, так и в разных. И это определяет различие сочетаемости.

Отрицательная поляризация изучалась, прежде всего, на примере англ. *any*. Было выявлено два значения *any*, NPI-*any*, как в (1), и **free choice**-*any* (сокр. FC-*any*), как в (2). В русском языке это будет, примерно, различие между *какой-нибудь* и *любой*:

(1) John didn't see *any* students there;
(2) *Any* cat hunts mice.

В некоторых контекстах это различие может быть передано противопоставлением кванторов ∀ и ∃ классической математической логики, пример из Haspelmath 1997: 95.

(3) If *anybody* can swim the channel then I can do it =
(i) Если *кто-нибудь* может переплыть этот канал, то и я могу;
(ii) Если *любой* может переплыть этот канал, то и я могу.

Однако если значение NPI-*any*, как и русского *какой-нибудь*, вполне удовлетворительно передается квантором ∃ (Падучева 1985: 94–98), то для FC-*any*, как и для русского *любой*, адекватного представления на языке логики нет (Haspelmath 1997: 95). Ключевым для отличия FC-*any* и *любой* от оператора ∀ является употребление в контексте иллокутивной модальности разрешения, когда контекст отводит место для субъекта выбора, который предусмотрен в семантике *любой*:

(4) You can take *any* apple ‘Можешь взять *любое* яблоко <по своему выбору>’.

Ниже речь идет о русских местоимениях с отрицательной поляризацией (ОП). Предлагается задать перечень ОП-контекстов списком и работать со списком.

4. Местоимения с отрицательной поляризацией в русском языке

Я исхожу из следующей классификации русских неопределенных местоимений (Падучева 1985, Haspelmath 1997).

I. **Референтные** неопределенные местоимения — допустимы в веридикативных контекстах.
1. Слабоопределенные (*Кое-кто тогда ко мне пришел*)
2. Неизвестности (*Кто-то пришел*)

II. **Нереферентные** неопределенные местоимения. Допустимы только в контексте операторов, снимающих веридикативность.
1. Отрицательные местоимения (*Никто ко мне тогда не пришел*)

2. Собственно нереферентные неопределенные местоимения, non-specific indefinite (*Возможно, я что-нибудь упустил*); в том числе — так наз. минимизаторы (*Можешь оказать хоть какую-нибудь помощь?*)
3. Отрицательно поляризованные (*Я не обязан отчитываться перед кем-либо /перед кем бы то ни было*)
4. Местоимения свободного выбора (*Возьми любую книгу; Спроси кого угодно*).

Отрицательное местоимение, т. е. местоимение на *ни*, — это, с семантической точки зрения, неопределенное местоимение в сфере действия отрицания (*никто* = 'неверно, что хоть кто-нибудь'), и оно может быть только нереферентным. При этом референтные неопределенные местоимения допустимы в отрицательном предложении (не в контексте вопроса или императива, см. Кобозева 1981, Падучева 1985, Haspelmath 1997: 40, 43) — поскольку они не входят в сферу действия отрицания:

(1) Я *чего-то* в твоем докладе не понял ≠ 'ничего';
(2) Он не смог *чего-то* увидеть [чего хотел] ≠ 'ничего'.

Референтные неопределенные местоимения не могут находиться в сфере действия оператора, снимающего веридикативность, например, модальности или условия. Так, в (3) *чего-то* = *чего-нибудь*:

(3) Если я *чего-то* не пойму, ты мне объяснишь.

В русском языке действует правило обязательного отрицательного согласования (*Никто мне тогда не помог*): местоимение на *ни-* требует отрицания при глаголе — аналогичное правило есть, например, в литовском, грузинском, греческом. В Giannakidou 2011 отрицательные местоимения в этих языках отнесены к классу NPI — к особому подклассу **строгих** NPI, к которому относятся англ. *either* и *yet*. Однако русские местоимения на *ни-* не относятся к NPI (вопреки Pereltsweig 2000). Прототипические отрицательно поляризованные единицы, такие как англ. *any* или русское *пальцем (не) пошевелит, чтобы…*, сами по себе отрицательными не являются. Между тем отрицательные местоимения сами являются отрицательными. Кроме того, все классические NPI возможны также и за пределами отрицательного контекста (англ. *either* и *yet* — очевидные изоляты), а русские отрицательные местоимения употребляются только в отрицательном контексте.

Русские NPI (иначе — ОП) рассматривались в Perelzveig 2000, Богуславский 2001, Татевосов 2002, Рожнова 2009, Падучева 2011. Задача данной работы — обосновать существование местоимений ОП как отдельного класса нереферентных местоимений русского языка; это местоимения на *-либо* и на *бы то ни было*. На базе Перечня контекстов будет предложено обоснование деления русских нереферентных местоимений на четыре класса: 1) отрицательные, 2) собственно нереферентные, на *нибудь*, 3) отрицательно поляризованные, на *-либо* и на *бы то ни было*, и 4) свободного выбора — *любой* и серия на *угодно*.

Перечень контекстов составлен на базе Haspelmath 1997, Падучева 1985, 2004, 2011, Giannakidou 2011 и др. Контексты сгруппированы так, чтобы обеспечить

возможность сопоставления контекстов серии на *-либо* и *бы то ни было*, с одной стороны, с контекстами, которые согласно Giannakidou 2002 лицензируют англ. NPI-*any*, а с другой — с контекстами других русских нереферентных неопределенных местоимений.

В Giannakidou 2011 утверждается, что NPI лицензируются тремя контекстами: отрицание, DE-операторы и неверидикативность. Однако отрицание — это просто один из неверидикативных контекстов. Ниже будет показано, что контекст DE-оператора, там, где он существен, можно свести к контексту неверидикативности. Тем самым все контексты употребления русских NPI будут представлены как контексты неверидикативности. Дело, однако, в том, контекст неверидикативности приютил не только NPI, но и другие классы нереферентных местоимений. Будет показано, что каждому классу соответствует свое подмножество контекстов из Перечня. (Контексты у *-либо* и *бы то ни было* совпадают не полностью, но различия минимальные.)

Условные обозначения. Отрицательные местоимения обозначаются как *ни*, NSI — как *нибудь*, местоимения серии на *-либо* и на *бы то ни было* — как ОП, местоимение *любой* и серия на *угодно* — как FC. Знак + означает допустимость ОП в данном контексте.

\+ 1. **Отрицание**

\+ а) Сопредикатное отрицание. Исключены *нибудь*, поскольку в контексте сопредикатного отрицания должно быть нереферентное отрицательное местоимение, на *ни*. Возможно ОП.

ни: Джон не видел там *никаких* студентов

**нибудь*: *Джон не видел там *каких-нибудь* студентов

ОП: Джон не видел там *каких бы то ни было /каких-либо* студентов

any: John didn't see *any* students there

\+ б) При отрицании в подчиняющей пропозиции, т. е. при «высоком» отрицании, *ни* возможны иногда, *нибудь* невозможны, возможны ОП.

ни: Он не хочет ничего менять; *Он не был расположен выслушивать никакие оправдания

*нибудь: *Он не был расположен выслушивать какие-нибудь оправдания

ОП: Он не был расположен выслушивать какие бы то ни было оправдания

any: John didn't want to do *anything*

\+ в) В контексте внутрисловного отрицания исключены местоимения *ни* и *нибудь*:

нибудь: *Бесполезно говорить им *что-нибудь*

Отрицательные местоимения невозможны нигде, кроме пп. а), б), и далее не упоминаются, а ОП допустимы — в контексте широкого круга слов с внутрисловным отрицанием: *отсутствовать, быть лишенным, отрицать, воздерживаться, исключать*; *прекратить, утратить, отнять, запретить, отменить, избавить, отказаться, отвергать, избегать, потерять, упразднить* и др. Например:

ОП: отсутствует *какая бы то ни было/какая-либо* внутренняя цензура [NB: возможно *всякая* в значении ‘даже малейшая’].

\+ 2. **Альтернатива**

\+ а) Вопрос, в том числе — косвенный; возможны *нибудь*, ОП и NPI-*any*.

нибудь: Он задавал тебе какие-нибудь вопросы?

ОП: Задавал ли он тебе какие бы то ни было /какие либо каверзные вопросы?

*FC: Он задавал тебе *любые /*какие угодно вопросы?

any: Did you see *anything*? He asked me whether I saw *anybody*

\+ б) Условие — контекст, где возможны местоимения всех трех типов.

нибудь: Если возникнут *какие-нибудь* проблемы, звони

ОП: Причём в случае, если возникнут какие бы то ни было /какие-либо экономические проблемы, виноват будет Кудрин; Решая какую бы то ни было частную задачу, надо думать о языке в целом

FC: Если возникнут любые проблемы, звони; Решая *любую* задачу, надо думать о языке

any: If you tell *anybody* about it, I’ll be upset

Контекст условия допускает FC, которые невозможны в контексте вопроса. При этом деепричастный оборот допускает FC шире, чем придаточное условное — поскольку он обеспечивает доступ к подразумеваемому субъекту выбора.

В контексте обособления может сниматься различие между ОП и FC: *как бы то ни было ≈ в любом случае ≈ так или иначе*:

безрелигиозный максимализм, *в какой бы то ни было /в любой* форме, ведет к деградации общества; *Каков бы ни был режим* — Россия наша Родина. [«Завтра», 2003.08.22]

Напротив, *нибудь* контексте обособления невозможно:

Сопротивление, в **какой-нибудь* форме, бесполезно.

– в) Дизъюнкция, т. е. контекст разделительного *или*; ОП возможны только при наличии вышестоящего отрицания — без отрицания здесь было бы местоимение на *нибудь* или даже референтное неопределенное.

ОП: Там нет ни трудностей, ни каких бы то ни было /каких-либо интересных задач

нибудь: Там есть трудности или какие-нибудь /какие-то интересные задачи

any: Either *anybody* came in or we left the light on

+/– 3. **Квантификация**

\+ а) Контекст квантора общности

В Giannakidou 2011 отрицательная поляризованность контекста универсальной квантификации, которая лицензирует *any*, объясняется ссылкой

на DE. На самом деле, определение в составе ИГ с универсальной квантификацией — это скрытая импликация, так что ОП в русском лицензируется в этом контексте совершенно так же, как в контексте условия:

ОП: Я всю косметику с *каким бы то ни было* ароматом убрала; Все студенты, которые знали *что бы то ни было /что-либо* о преступнике, вступили в контакт с полицией.
any: Every student who saw *anything* should report to the police

Если *нибудь* употребляется вне подчинительного контекста, оно выражает **дистрибутивность** — каждый знал что-то свое:

нибудь: Все студенты знали *что-нибудь* о преступнике.

+ б) Контекст кванторов, которые являются DE-операторами:

Не требуется обращения к DE также для того, чтобы объяснить допустимость ОП в контексте *мало*, но не в контексте *много*: *мало* содержит внутрисловное отрицание.

ОП: Мало кто имел *какое бы то ни было* представление о предмете; ср. *Многие имели *какое бы то ни было* представление о предмете
any: Few children saw *anything*, ср. *Many professors invited *any* students

ОП возможны в контексте числового квантора существования. Источник неверидикативности этого контекста обсуждается в разделе 5.

ОП: Ровно три человека видели что бы то ни было; ОП: Самое большее пять человек видели что бы то ни было /*что-нибудь
any: Exactly three students saw *anything*; At most five people saw *anything*

–/+ 4. **Узуальность и многократность**

Внутрисловное отрицание объясняет допустимость ОП в контексте *редко*; а *часто, обычно, иногда* не лицензируют ОП; *нибудь* возможно в значении дистрибутивности:

– а) в контексте *часто, обычно, иногда*
нибудь: Он часто *что-нибудь* забывает; Дарья обычно звонит *кому-нибудь*, когда ей скучно
*ОП: Он часто **что-либо /*что бы то ни было* забывает; Дарья обычно звонит **кому-либо /*кому бы то ни было*, когда ей скучно
any: He usually reads *any* book very carefully.

+ б) в контексте *редко*:
нибудь: Он редко что-нибудь забывает;

ОП: Он редко обсуждает с кем-либо свои дела; Он редко забывает что бы то ни было.
any: He usually reads *any* book very carefully.

Дальше идут контексты Будущее время, Модальность и Желание, которые не лицензируют ОП: они оптимальны для *нибудь*.

– 5. **Будущее время**
нибудь: Джон купит какую-нибудь бутылку вина
*ОП: Джон купит *какую бы то ни было /*какую-либо бутылку вина
any: John will buy *any* bottle of wine.

– 6. **Модальность**
– а) возможность; ОП исключены; уместны *нибудь* и FC.
нибудь: Он может что-нибудь знать
*ОП: Джон может уговорить *кого-либо /*кого бы то ни было
FC: Джон может разозлить любого /кого угодно; Комитет может предоставить работу любому кандидату
any: John may persuade *anybody*; The committee can give the job to *any* candidate

У *любой* может возникать скалярное значение ‘даже наименее вероятный’. Вопреки Haspelmath 1997: 119, речь не идет о минимальной или максимальной точке.

– б) необходимость; ОП исключены, но и FC не вполне на месте; уместны *нибудь*.
нибудь: Вы должны принести какую-нибудь справку
*ОП: Вы должны принести *какую-либо /*какую бы то ни было справку
FC: Вы должны принести любую справку
any: Any minors must be accompanied by their parents

– 7. **Желание**
– а) просьба; ОП исключены; широко употребляются *нибудь*, в том числе — в значении ‘хоть’, ср. об *any* в значении минимизатора в Giannakidou 2011. FC исключены.
нибудь: Предложите мне что-нибудь взамен; Дайте хоть что-нибудь!
ОП: Предложите мне взамен *что бы то ни было (возможно что-либо)
FC: #Задумайте любое число [это не просьба]
any: John asked us to invite *any* student; I ask you to give me *any* apple
– б) разрешение, согласие, готовность; ОП исключены; возможны *нибудь*; широко употребляются FC. Контекст готовности: *готов заплатить любую сумму* = ‘даже самую большую’; но *готов на любую зарплату* =

‘даже на самую малую’. Только в контексте достаточного условия *любой* отсылает к минимальной точке.

нибудь: Разрешаю взять с собой что-нибудь теплое
*ОП: *Я согласен добавить что бы то ни было
FC: Я согласен что угодно добавить; Я готов на любую работу; достаточно любой мелочи
any: Take any apple; You may take any apple

– 8. **Пропозициональные установки: мнение**

– а) Глаголы мнения, хотя их подчиненная пропозиция неверидикативная, не лицензируют ОП; допустимо *нибудь*:

нибудь: Я надеюсь, что это кому-нибудь /*кому бы то ни было известно
*ОП: *Он боится каких бы то ни было репрессий
*any: *John believes that we invited *any* student

Более того, сослагательное наклонение не создает контекста для ОП:

Это решило бы *какие-нибудь /*какие бы то ни было* из наших проблем.

ОП возможны только при внутрисловном отрицании в подчиненной пропозиции:

Он боится *что бы то ни было* менять.

+ б) Неуверенность, предположительность, нереальность (в т. ч. англ. *hardly* и *barely*; русское *едва ли*) лицензируют ОП:

ОП: Бесполезно говорить им что бы то ни было; Едва ли он учился чему бы то ни было
any: It was useless to tell him anything; John hardly talked to anybody; John barely studied *anything*

В контексте глагола с внутрисловным отрицанием *нибудь* и ОП синонимичны:

нибудь: Сомневаюсь, что это кому-нибудь известно
ОП: Сомневаюсь, что это известно кому бы то ни было.

– 9. **Сравнение**

В Giannakidou 2009 утверждается, что контекст сравнения не может быть охарактеризован ни как DE, ни как отрицательный, и поэтому в контексте сравнительного оборота не NPI-*any*, а FC-*any*. Перевод на русский — *любой*.

any: He caused more harm than any terrorist ‘он нанес больше вреда, чем любой террорист’.
ОП: В любом случае ей с ребёнком лучше находиться дома, чем *где бы то ни было*. [Л. Улицкая. Путешествие в седьмую сторону света] = ‘чем в *любом* другом месте’.

Поэтому проведение в жизнь этих директив сейчас более необходимо, чем *когда бы то ни было /когда-либо* = ‘чем в *любое* другое время’.

+ 10. **Временной предлог *прежде чем* и *как только***

Предлог *прежде чем* может порождать неверидикативный контекст и лицензировать ОП (ср. Zwarts 1995); не исключены и *нибудь*:

ОП:	Он ушел прежде, чем я успел что-либо сказать; Она позвонит, как только узнает что-либо /что-нибудь
any:	He left before I could say *anything*

То же в сравнительных оборотах:

Он скорее удавится, чем скажет что-нибудь приятное.

В семантику предлогов *без, вне, вместо* входит внутрисловное отрицание; подчиненная пропозиция попадает в его сферу действия, так что они не добавляют принципиально новых контекстов к тем, что были рассмотрены. Внутрисловное отрицание в контексте *без* — это стилизованная аномалия:

ни:	А Зинка, как ненормальная, отдала <деньги> щербатой без *никакой* расписки, на веру. [И. Грекова. Перелом (1987)]

5. Обсуждение результатов

Итак, местоимения на *либо* и *бы то ни было* обычно укладываются в контекст неверидикативности. Пропозиция в контексте отрицания всегда неверидикативная. Поэтому найти отрицание — значит найти неверидикативный контекст.

Важно, что на возможность употребления ОП влияет не поверхностное отрицание , а именно отрицательный смысл. Так, в примерах (1а), (2а) внутрисловное отрицание лицензирует ОП, а в (1б), (2б) двойное отрицание дает положительный смысл, так что ОП исключено:

(1) а. Он врет, что читал *что-либо /что бы то ни было*;
б. Он не врет, что читал **что-либо /*что бы то ни было*.

(2) а. Способность этого человека к *какой бы то ни было* деятельности не очевидна;
б. Способность этого человека к **какой бы то ни было /*какой-либо* деятельности не вызывает сомнений.

Пример (3) выявляет релевантный отрицательный компонент в составе слова *конец* — в контексте *начало* ОП неуместно:

(3) а. Это конец *какой бы то ни было* свободной экономике;
б. *Это начало *какой бы то ни было* свободной экономики.

Семантическое разложение слова *только* выявляет в нем компонент отрицание. И действительно, *только* лицензирует *какой бы то ни было* (ср. Giannnakidou 2006):

(4) а. Только Петя понял *что бы то ни было* в этом докладе
б. *Петя понял *что бы то ни было* в этом докладе.

Отрицание может входить в значение конструкции. Конструкция *слишком …, чтобы* содержит подспудное отрицание:

Я слишком устал, чтобы спорить *с кем бы то ни было* ⊃ ‘не стану спорить’.
Ребенок слишком мал, чтобы понимать *что бы то ни было* ⊃ ‘не понимает <ничего>’.

Отрицание в вопросе задает утвердительное предположение, которое и отпугивает *какой бы то ни было*:

Задавал ли он тебе *какие-нибудь /какие бы то ни было* вопросы?
Не задавал ли он тебе *каких-нибудь /*каких бы то ни было* вопросов?

В примере (5), по мотивам Giannnakidou 2006, допустимо *какой бы то ни было*, хотя фактивные глаголы не порождают невериликативности в пропозициональном актанте. Видимо, ОП лицензируется отрицанием в семантике *сожалеть*:

(5) Он *сожалеет*, что принимал от нее *какие бы то ни было* подарки [было бы лучше, если бы не принимал никаких].

В примере (6) (на базе Рожнова 2009) речь идет о ситуации, которая имела место ровно один раз. Откуда берется невериликативность? А дело в том, что пропозиция входит в сферу действия числового квантора существования — утверждается не то, что произошло событие P, а то, что событие, состоящее в том, что P, произошло ровно один раз:

(6) Это был единственный раз, когда я *кого бы то ни было* ударил.

Аналогично объясняются ОП в контексте числового квантора существования в разделе 4.

В примере (7) (на базе Giannakidou 2006) тоже единичная ситуация. Тут минимизатор: *какой бы то ни было* = ‘хоть какой-нибудь’:

(7) Он был до смерти рад, что мы достали ему *какой бы то ни было* билет.

Минимизатор *хоть* сочетается с *нибудь*, но не с ОП. В примере ниже *хоть* сочетается с *что-либо*, поскольку оно употреблено в значении ‘что-нибудь’; *хоть* с *бы то ни было* не сочетается вообще:

Пусть кто-нибудь укажет *хоть что-либо* подобное. [Н. И. Бухарин]

6. Заключение

Итак, обосновано существование в русском языке разряда местоимений с отрицательной поляризацией, к которым относится две серии местоимений — на *-либо* и на *бы то ни было*. Местоимения этих серий употребляются в составе пропозиции, которая находится в сфере действия отрицания, а также в ряде других контекстов, как это свойственно прототипическим NPI в разных языках: таких, как вопрос или протасис условного предложения.

На русском материале обосновано положение, которое отстаивается в ряде работ по формальной семантике: условием, ограничивающим употребление NPI, является, прежде всего, неверидикативный контекст — контекст пропозиции, которая употребляется безотносительно к истине. Неверидикативный контекст был ранее выявлен как контекст, составляющий условие употребления для русских нереферентных местоимений на *-нибудь*. Местоимения на *-нибудь* задают максимально широкий класс контекстов нереферентности. В разделе 4 приводится перечень этих контекстов. Из них для NPI исключен контекст прямого (т. е. сопредикатного) отрицания, поскольку в этом контексте обязательно должно быть употреблено отрицательное местоимение, а также контексты, в которых уместно NSI и FC. На долю NPI остаются контексты, в которых так или иначе прослеживается отрицание. Это Альтернатива (Вопрос и Условие), Сомнение.

Русские ОП не употребляются в контекстах Будущее, Модальность, Желание, в том числе — Императив, Дизъюнкция, Нефактивные пропозициональные установки. Все это контексты для серии *нибудь*, т. е. для NSI. Дистрибутивность в контексте универсальности, многократности и узуальности тоже обслуживают NSI, а не ОП. То, что они лицензируют англ. *any*, может означать, что у *any*, помимо двух общепризнанных значений, NPI и FC, есть значение non-specific indefiniteness.

Теория речевых актов и модальность внесли ясность в семантику местоимений свободного выбора — загадочного русского *любой*. FC тоже изымает из сферы ОП некоторые контексты: FC, а не ОП, обслуживают контексты Возможность, Генеричность, Сравнение.

Сравнение понятий неверидикативность и снятая утвердительность имело двоякий эффект. С одной стороны, неверидикативность позволила уточнить понятие снятой утвердительности как безотносительность к истине. С другой стороны, это сравнение позволило соотнести неверидикативность с полезным понятием, которое возникло в американской лингвистике еще в 60-е годы прошлого века, но было забыто.

Литература

1. *Boguslavskij I. M.* (2001) Modality, comparativeness and negation [Modal'nost', sravnitel'nost' i otricanie], Russian language in a scientific light [Russkij jazyk v nauchnom osveshchenii], Vol. 1. pp. 27–50.
2. *Boguslavskij I. M.* (2008) Between truth and falsity: adverbials in the context of suspended assertion [Mezhdu istinoj i lozh'ju: adverbialy v kontekste snjatoj utverditel'nosti], Logical analysis of language: between falsity and fantasy [Logicheskij analiz jazyka: mezhdu lozh'ju i fantaziej], Indrik, Moscow, 67–277.
3. *Borshchev V. B., Paducheva E. V., Partee B., Testelec Ja. G., Janovich I. S.* (2008) Russian Genitive case: referentiality and semantic types [Russkij roditel'nyj padezh: referentnost' I semanticheskie tipy] In Object genitive of negation in Russian. Researches in the theory of grammar [Objektnyj genitiv pri otricanii v russkom jazyke. Issledovanija po teorii grammatiki]. Vol.5, Probel-2000, Moscow, 148–175.
4. *Church A.* (1960) Introduction to Mathematical Logic [Vvedenie v matematicheskuju logiku], Translation into Russian, Inostrannaja literatura, Moscow.
5. *Giannakidou A.* (1998) Polarity sensitivity as (non)veridical dependency. John Benjamins, Amsterdam.
6. *Giannakidou A.* (2001) The meaning of free choice (https://webshare.uchicago.edu/users/giannaki/Public/pubs/GiannaLP2001.pdf), Linguistics and Philosophy, Vol. 24, pp. 659–735.
7. *Giannakidou A.* (2006) *Only*, emotive factives, and the dual nature of polarity dependency, Language, Vol. 82, pp. 575–603.
8. *Giannakidou A.* (2011). Positive polarity items and negative polarity items: variation, licensing, and compositionality (https://webshare.uchicago.edu/users/giannaki/Public/pubs/HSK.chapter64.proofs.pdf). In Semantics: An International Handbook of Natural Language Meaning. Berlin: Mouton de Gruyter, pp. 1660–1712.
9. *Haspelmath M.* (1997) Indefinite Pronouns, Clarendon Press, Oxford.
10. *Karttunen L., Zaenen A.* (2005) Veridicity, in Annotating, Extracting and Reasoning about Time and Events (http://drops.dagstuhl.de/portals/index.php?semnr=05151). Dagstuhl Seminar Proceedings 05151, G. Katz, J. Pustejovsky, F. Schilder (eds.), Dagstuhl, Germany, 2005.
11. *Klima E.* (1964) Negation in English, in Janet Fodor and Jerrold Katz (eds.) The structure of language, Prentice-Hall Inc: Englewood Cliffs, pp. 246–323.
12. *Kobozeva I. M.* (1981) Towards pragmatic analysis of the Russian *to-* and *nibud'*-pronouns [Opyt pragmaticheskogo analiza *to-* i *nibud'*-mestoimenij]. Proceedings of the Russian Academy of Sciences, Series of Language and Literature [Izvestija AN SSSR, Serija literatury i jazyka], Vol. 40, 165–172.
13. *Ladusaw W. A.* (1980) Polarity Sensitivity as Inherent Scope Relations, Ph.D. dissertation, published by New York: Garland Publishing.
14. *Paducheva E. V.* (1985) Utterance and its relationships with reality [Vyskazyvanie i ego sootnesennost' s dejstvitel'nost'ju], Nauka, Moscow; 6-th edition 2010, Izdatel'stvo LKI, Moscow. http://lexicograph.ruslang.ru/TextPdf1/paducheva1985.pdf

15. *Paducheva E. V.* (2004) Effect of suspended assertion [Effekt snjatoj utverditel'nosti], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2004" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2004"], Verhnevolzhskij, pp. 479–486. http://lexicograph.ruslang.ru/TextPdf2/dialog_2004_Paducheva.pdf.
16. *Paducheva E. V.* (2005) Effects of suspended assertion: global negation [Effekty snjatoj utverditel'nosti: global'noe otricanie] Russian language in a scientific light [Russkij jazyk v nauchnom osveshchenii], Vol. 2(10), pp. 17–42. http://lexicograph.ruslang.ru/TextPdf2/ryns2005.pdf
17. *Paducheva E. V.* (2011) Implicit negation and negative polarity pronouns [Implicitnoe otricanie i mestoimenija s otricatel'noj poljarizaciej], Problems of linguistics [Voprosy jazykoznanija], Vol.1, pp. 3–18. http://lexicograph.ruslang.ru/TextPdf1/vnutrilex_neg-VJa.pdf
18. *Rozhnova M. A.* (2009) Syntactic properties of negative pronouns in Spanish and Russian [Sintaksicheskie svojstva otricatel'nyh mestoimenij v ispanskom i russkom jazykax], Diploma paper, RGGU, Moscow.
19. *Tatevosov S. G.* (2002) Semanics of constituents of a noun phrase: quantifier words [Semanika sostavljaushchih imennoj gruppy: quantornye slova] IMLI RAN, Moscow.
20. *Pereltswaig A.* (2000) Monotonicity-based vs. veridicality-based approaches to negative polarity: evidence from Russian. in Tracy Holloway King and Irina A. Sekerina (Eds.) Formal Approaches to Slavic Linguistics: The Philadelphia Meeting 1999. pp. 328–346. Ann Arbor: Michigan Slavic Publications.
21. *Vendler Z.* (1967)Causal relations, in Journal of philosophy, 1967, Vol. 64, pp. 704–713.
22. *Zwarts F.* (1995) Nonveridical Contexts, Linguistic Analysis, Vol. 25, pp. 286–312.

ИНДЕКС ТОНАЛЬНОСТИ РУССКОЯЗЫЧНОГО ФЕЙСБУКА

Панченко А. И. (alexander.panchenko@uclouvain.be)

Лувенский католический университет, Лувен, Бельгия;
ООО «Лаборатория Цифрового Общества», Москва, Россия

Индекс тональности измеряет эмоциональный уровень в корпусе текстов. В данной работе мы определяем четыре подобных индекса. Предложенные индексы используются для измерения общего уровня «позитивности» группы пользователей на основании их постов в социальной сети. В отличие от предыдущих работ, мы впервые вводим индексы, которые работают на уровне текстов, а не отдельных слов. Кроме этого, данная работа впервые представляет результаты вычисления индекса тональности на значительной части русскоязычного Фейсбука. Полученные результаты согласуются с результатами подобных экспериментов на англоязычных корпусах.

Ключевые слова: анализ тональности текста, индекс тональности, Фейсбук

SENTIMENT INDEX OF THE RUSSIAN SPEAKING FACEBOOK

Panchenko A. I. (alexander.panchenko@uclouvain.be)

Universite catholique de Louvain, Louvain-la-Neuve, Belgium;
Digital Society Laboratory LLC, Moscow, Russia

A sentiment index measures the average emotional level in a corpus. We introduce four such indexes and use them to gauge average “positiveness” of a population during some period based on posts in a social network. This article for the first time presents a text-, rather than word-based sentiment index. Furthermore, this study presents the first large-scale study of the sentiment index of the Russian-speaking Facebook. Our results are consistent with the prior experiments for English language.

Keywords: sentiment analysis, sentiment index, Facebook, Gross National Happiness

1. Introduction

Social media analysis has opened new exciting possibilities for social and economic sciences [14]. In this paper, we present a technique for measuring *social sentiment index*. This metric reflects the emotional level of a social group by means of text analysis. As Dodds et al. [5] notice, "a measure of societal happiness is a crucial adjunct to traditional economic measures such as gross domestic product".

Social sentiment index is based on the sentiment analysis technology. There is a huge number of papers on sentiment analysis, most notable of English [1,6,7,8], but also of Russian [9, 10] texts. However, most of these studies focus on classification of a single text that expresses sentiment about a particular entity. Some attempts were also made to calculate sentiment indexes. For instance, Godbole et al. [1] performed sentiment analysis of news and blogs and computed a sentiment index of named entities, such as George Clooney or Slobodan Milosevic.

Several researchers tried to measure "positiveness" or "happiness" of social network users. Most experiments with the happiness indexes were conducted on English texts. Several works dealt with the emotionally-annotated posts from LiveJournal, the biggest online community back to 2005. Mihalchea and Lui [12] analyzed some 10,000 posts from LiveJournal, annotated with the tags "happy" or "sad". The authors identified terms that characterize happiest and saddest texts. According to this study, the happiest day is Saturday, while the happiest hours are 3 am and 9 pm. Mishne and Rijke [17] performed another experiment on the LiveJournal data. The authors collected a corpus of 8 million posts labeled with 132 distinct mood tags such as "sad", "drunk" or "happy". This dataset was used to train supervised models predicting tags by texts. In particular, this work studied feedback of the blogosphere on a terror attack in London. Balog and Rijke [18] performed a time series analysis of 20 million LiveJournal posts issued during one year (2005–2006). The researchers observed a clear impact of the weather, holidays and season on the mood of some profiles.

More recent studies are focused on analysis of social networks, such as Twitter and Facebook. Dodds et al. [5] analyzed 4.6 billion of tweets over a period from 2008 to 2011. According to the authors, the dataset represented 5% of Twitter back to 2011. The researchers measured "happiness" with a simple dictionary-based approach and came to several conclusions: the happiest day of the week is Saturday; the happiest hour is 5 am; peaks of happiness coincide with national holidays, such as Christmas and St. Valentine's day.

Kamvar and Harris [15] developed a system that maintains an index of 14 million phrases of 2.5 million people starting from "I feel" or "I am feeling". The phrases along with information about their respective authors were crawled from blogs, microblogs and social networks. The system is able to answer questions like "Do Europeans feel sad more often than Americans?" or "How did young people in Ohio feel when Obama was elected?".

The work of Kramer [13] is probably the most similar to our experiment. The author analyzes use of emotion words by roughly 100 million Facebook users since 2007. He proposed a "Gross National Happiness" index, based on a standardized difference between positive and negative terms in texts. The peaks of this index coincide with the nationally important dates.

Recent work of Wang et al. [16] challenges the idea of the Gross National Happiness index by comparing its results to the data gathered with a Facebook user questionnaire "myPersonality"[1]. The authors did not find a significant correlation between self-reported happiness and happiness computed from their posts.

Two contributions of this paper to the exploration of sentiment/happiness indexes are as follows:

1. We propose four indexes that gauge emotional level of population. Unlike most previous works, we not only deal with word-based indexes, but also propose a text-based sentiment index.
2. To the best of our knowledge, this is the first study of sentiment index of the Russian-speaking social network.

2. Social Sentiment Index

First, this section presents several network datasets used to calculate social sentiment indexes. Next, we describe sentiment analysis method that is used in calculations. Finally, we present four indexes used in our experiment.

2.1. The Facebook Corpus

We use in our experiments a corpus of Facebook posts provided for research purposes by Digsolab LLC[2]. This anonymized dataset contains texts from the publically available part of the social network. According to our collaborator at Digsolab, the corpus was collected using the API of Facebook[3]. Construction of such anonymized samples from Facebook is common in the research community [12]. The dataset we deal with contains 573 million anonymized posts and comments of 3,190,813 users. We considered only texts in Russian. The language detection was completed using the *langid.py* module [2]. Table 1 summarizes key parameters of the dataset.

The oldest post in the corpus dates back to the 5th August of 2006, while the latest is of 13th November 2013. The distribution of posts over time is far from being uniform (see Fig. 1). As one may expect, the number of posts in the social network grows rapidly as the network gets popular with Russian-speaking users. The biggest number of texts per day (1,243,310) was observed on the 4-th April 2013. After this peak, the number of posts drops to 600–700 thousand per day. This is an artifact of the data collection method used. However, as sentiment index is a relative value (see Section 2.3) it should not be much affected by such noise.

---

[1] http://www.psychometrics.cam.ac.uk/productsservices/mypersonality

[2] http://www.digsolab.com/

[3] https://developers.facebook.com/tools/explorer?method=GET&path=EuropeanCommission%2Fposts

Table 1. Statistics of the Facebook corpus

Number of anonymized users	**3,190,813**
Language	Russian
Number of posts	426,089,762
Number of comments	147,140,265
Number of texts (posts + comments)	**573,230,027**
Number of tokens in posts	20,775,837,467
Number of tokens in comments	2,759,777,659
Number of tokens (posts + comments)	**23,535,615,126**
Average post length, tokens	49
Average comment length, tokens	19

Facebook claims [3] that by March 2013 it had about 1.19 billion active users. According to Internet World Stat [4], the number of Facebook users in Russia by the end of 2012 is 7,963,400. Thus, our sample is roughly equal to 40% of the 2012 Russian Facebook.

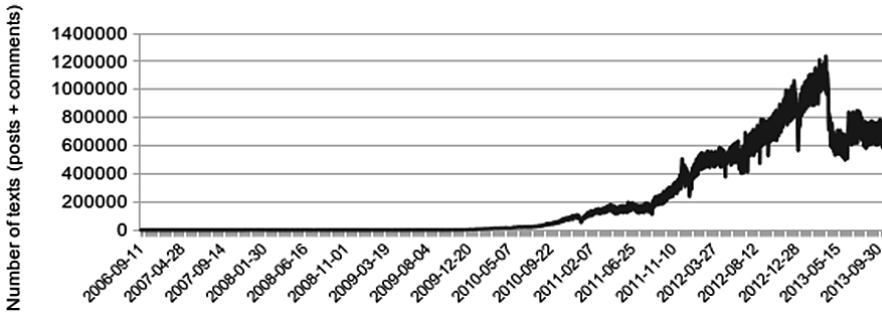


Fig. 1. Number of texts per day in the Facebook corpus

2.2. Sentiment Detection Method

The social sentiment index builds upon a popular approach to sentiment analysis, based on a dictionary of positive and negative terms [6,7]. Similar approach was employed by Dodds et al. [5] to measure "happiness" of Twitter users.

We use a custom sentiment dictionary developed specifically for the social media texts. Two annotators independently labeled 15,000 most frequent adjectives in the aforementioned corpus. The result is a dictionary of 1,511 terms, where each term is supposed to be a context-independent predictor of a positive/negative opinion. Each term was labeled by both annotators as positive or negative. This dictionary contains 600 negative terms (D_-) and 911 negative terms (D_+) such that each term $w \in \{D_- \cup D_+\}$ was labeled as positive or negative by both annotators. Table 2 presents some examples of these terms.

Table 2. Most frequent positive and negative adjectives in the Facebook corpus

Positive adjectives		Negative adjectives	
хороший	good	плохой	bad
новый	new	старый	old
первый	first	долгий	long
нужный	helpful	неблагоприятный	unfavorable
бесплатный	free of charge	скучный	boring
любимый	beloved	сложный	complicated
интересный	interesting	голодный	hungry
спокойный	quiet	страшный	scary
социальный	social	скучно	bored
добрый	kind	немой	mute

Each text t is represented as a multiset of n lemmas $\{w_1, \ldots, w_n\}$. The lemmatization was done with the morphological analyzer *PyMorphy*[4]. We used the following decision rule to classify a text t as positive (+1), negative (−1) or neutral (0):

$$c(t) = \begin{cases} +1 & if \quad |\Delta| \geq \alpha \ \ and \ \ \Delta > 0 \\ 0 & if \quad |\Delta| < \alpha \\ -1 & if \quad |\Delta| \geq \alpha \ \ and \ \ \Delta < 0 \end{cases}, \ where \ \Delta = \frac{|\{w \in t \mid w \in D_+\}| - |\{w \in t \mid w \in D_-\}|}{n}$$

Here the *emotion delta* $\Delta \in [-1; 1]$ shows both direction and amplitude of a text sentiment. According to the decision rule mentioned above, a text is classified as neutral if Δ is less than some constant α. In our experiments, we fixed $\alpha = 0.05$. This value provides a good tradeoff between precision and recall (see Table 3).

We deliberately avoided machine learning, as Facebook is an open domain corpus. Statistical sentiment classifiers trained on one domain are known to perform poorly on another domain [6,7], as such classifiers can heavily rely on terms that are positive or negative only within a specific field. According to Pang and Lee [6], "simply applying the classifier learned on data from one domain barely outperforms the baseline for another domain". Therefore, we rather opted for a dictionary-based approach, that relies on a set of *frequent domain-invariant* positive and negative terms.

We compared the dictionary-based classification method with the baselines on the ROMIP 2012 dataset [10]. This dataset contains some 50 thousands of reviews about films, movies and digital cameras. Results of this evaluation are presented in Table 3.

[4] https://bitbucket.org/kmike/pymorphy

Table 3. Performance of the dictionary-based sentiment classification approach, as compared to other methods (ROMIP-2012 dataset)

RunID	Object	Macro_P	Macro_R	Macro_ F1	Accu- racy	P_1	P_0	P_-1	R_1	R_0	R_-1
xxx	books	0.379	0.443	0.350	0.536	0.873	0.157	0.106	0.604	0.157	0.569
sentistrength	books	0.368	0.403	0.327	0.448	0.848	0.154	0.103	0.468	0.375	0.367
yyy	books	0.399	0.494	0.377	0.560	0.908	0.154	0.136	0.620	0.183	0.678
nb-blinov	books	0.408	**0.528**	**0.390**	**0.675**	0.909	0.157	0.157	0.785	0.042	**0.757**
dict (α = 0.02)	books	0.42	0.431	0.348	0.445	0.893	**0.175**	0.191	0.42	0.677	0.197
dict (α = 0.05)	books	0.437	0.404	0.274	0.327	0.919	0.163	0.229	0.246	0.844	0.122
dict (α = 0.07)	books	**0.446**	0.381	0.217	0.261	**0.934**	0.157	0.248	0.155	**0.911**	0.077
Xxx	movies	0.395	0.454	0.361	0.493	0.819	0.235	0.131	0.586	0.148	0.628
Sentistrength	movies	0.371	0.401	0.343	0.436	0.774	0.219	0.119	0.485	0.274	0.445
Yyy	movies	0.411	**0.497**	0.390	0.522	0.849	0.221	0.165	0.610	0.173	**0.708**
nb-blinov	movies	0.347	0.489	**0.400**	**0.705**	0.788	0.000	0.253	**0.943**	0.000	0.525
dict (α = 0.02)	movies	0.453	0.438	0.383	0.465	0.852	0.264	0.241	0.416	0.723	0.177
dict (α = 0.05)	movies	0.473	0.412	0.315	0.382	0.892	**0.249**	0.278	0.259	0.869	0.109
dict (α = 0.07)	movies	**0.48**	0.388	0.26	0.329	**0.908**	0.239	**0.293**	0.172	**0.923**	0.069
Xxx	cameras	0.388	0.390	0.370	0.561	0.864	0.111	0.190	0.632	0.138	0.400
Sentistrength	cameras	0.373	0.359	0.319	0.429	0.855	0.106	0.157	0.461	0.370	0.247
Yyy	cameras	0.445	**0.488**	**0.443**	0.628	0.903	**0.127**	0.303	0.687	0.312	**0.464**
nb-blinov	cameras	0.445	0.441	0.437	**0.816**	0.844	0.000	**0.490**	0.974	0.000	0.350
dict (α = 0.02)	cameras	0.452	0.359	0.294	0.411	0.889	0.102	0.365	0.439	0.564	0.073
dict (α = 0.05)	cameras	0.441	0.329	0.193	0.256	0.905	0.095	0.324	0.232	0.733	0.022
dict (α = 0.07)	cameras	**0.462**	0.32	0.141	0.181	**0.912**	0.093	0.381	0.13	**0.816**	0.015

Our method is denoted as *dict*. The table also features performances of four alternative sentiment classification methods. The *xxx* and *yyy* are commercial sentiment classification systems[5]. The system *sentistrength* is a freely available tool for research purposes, developed by Thelwall et al. [8]. In this benchmark, we used standard Russian dictionaries coming with this tool with default parameters. The system *nb-blinov* is based on the sentiment phrase dictionary developed by Blinov et al. [9]. The dictionary specifies conditional probabilities of 19,000 phrases given a positive or negative class, e.g.:

$$p(w|-1) = p(\text{еле досматривать}|-1) = 0.000881168;$$
$$p(w|+1) = p(\text{еле досматривать}|+1) = 0.000016001.$$

The method *nb-blinov* relies on the decision rule, used in the Naïve Bayes classifier:

$$c(t) = \arg \max_{c \in \{-1,0,+1\}} p(c) \prod_{\{w \in T \mid w : D_+ \cup D_-\}} p(w|c)$$

Table 3 presents results of the comparison in terms of precision (Macro_P, P_-1, P_0, P_1), recall (Macro_R, R_-1, R_0, R_1), F-measure (Macro_F1) and accuracy.

[5] We cannot reveal their names, as the developers who provided the demo-versions did not let us do so. The *xxx* relies on both machine learning and rules, while *yyy* is a rule-based system.

As we may observe, the precision of the dictionary-based technique on the positive and negative class matches or even outperforms precision of the other systems. On the other hand, the recall of the *dict* method is significantly lower as compared to other techniques. As one may expect, the bigger the value of the parameter the higher the precision and the lower the recall.

We conclude that the dictionary-based method described in Section 2.2 is suitable for the needs of the social sentiment index, as it is domain-independent and yields the baseline performance.

2.3. Sentiment Indexes

The goal of a sentiment index is to measure positiveness of social network users during some period. For instance, one may want to know if high values of the index are related to national holidays, such as New Year, while low values of the index occur during "depressing" periods related to national tragedies, such as air crashes. Sentiment index can be useful (1) to analyze feedback of social media users on notable events and (2) for prediction of "planned" peaks of positiveness/negativeness. In our experiment, we used the following four simple metrics:

1. **Word Sentiment Index** is a ratio of positive to negative terms in all texts (posts and comments) in a corpus *T*:
$$s_w = \frac{\sum_{t \in T} |\{w \in t \mid w \in D_+\}| + \epsilon}{\sum_{t \in T} |\{w \in t \mid w \in D_-\}| + \epsilon}$$
2. **Text Sentiment Index** is a ratio of positive to negative texts in the corpus *T*:
$$s_t = \frac{|\{t \in T \mid c(t) = +1\}| + \epsilon}{|\{t \in T \mid c(t) = -1\}| + \epsilon}$$
3. **Word Emotion Index** is a ratio of emotional (positive or negative) words in texts in the corpus *T*:
$$e_w = \frac{\sum_{t \in T} |\{w \in t \mid w \in \{D_+ \cup D_-\}\}| + \epsilon}{\sum_{t \in T} |t| + \epsilon}$$
4. **Text Emotion Index** is a ratio of emotional (positive or negative) texts in the corpus *T*:
$$e_t = \frac{|\{t \in T \mid c(t) = \{-1, +1\}\}| + \epsilon}{|T| + \epsilon}$$

Both text and word sentiment indexes are in the range [0; $+\infty$]; here $\in$ was set to 10^{-6}. The first index works on the lexical level, while the second index deals with texts. Thus, precision of the second index depends on precision of the decision rule *c(t)*. Finally, the last two indexes capture the overall emotional level in a social network.

3. Results and Discussion

Table 4 presents results of the sentiment analysis of the entire corpus. According to both word-, and text-based indexes, positiveness predominate over negativeness. Users employ positive terms approximately 3.8 times more often than the negative terms. People use more emotional terms in the comments and fewer positive or negative terms in posts. Our results show that, during the entire period, people issued roughly 7.5 times more positively oriented texts than negatively oriented texts. Thus, we see that users of the Facebook tend to write much more positive texts, than negative ones. This trend is common for both posts and comments.

Indeed, people are keen to share happy moments and are reluctant to share dramas in their lives. Furthermore, by analyzing only public data, we probably miss a significant share of posts with a negative sentiment. One can expect that users tend to discuss negative matters in close circle of friends.

Table 4. Results of the social sentiment index calculation (the entire period)

	posts, %	**comments, %**	**posts + comments, %**
positive words	1.38	2.06	1.72
negative words	0.37	0.47	0.42
Word Emotion Index, e_w	1.75 (0.017)	2.53 (0.025)	2.14 (0.021)
Word Sentiment Index, s_w	3.72 (0.037)	4.38 (0.044)	3.81 (0.038)
positive texts	13.43	18.42	14.71
negative texts	1.83	2.28	1.94
Text Emotion Index, e_t	15.26 (0.153)	20.70	16.65 (0.166)
Text Sentiment Index, s_t	7.34 (0.073)	8.08	7.58 (0.076)

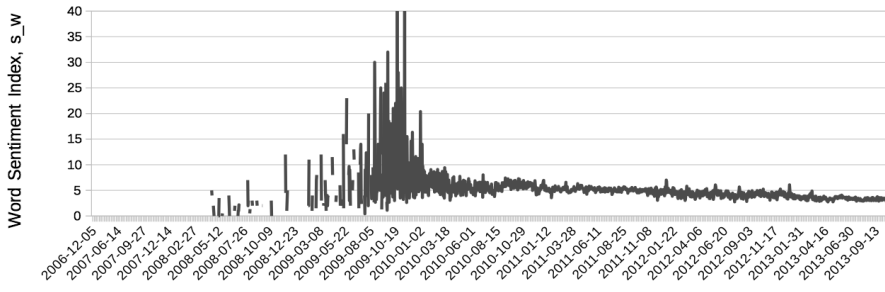


Fig. 2. Evolution of the word sentiment index s_w during the entire period

Fig. 2 shows dynamics of the word sentiment index s_w during the entire period. First, values of the index are significantly greater than one most of the time. Second, the index is fluctuating greatly from day to day. As we already learned from Fig. 1, most texts in the dataset were issued after the 1-st January of 2011. We suppose that, the huge fluctuations of the sentiment index from 2006 until the

end of 2010 are due to sparseness/incompleteness of the data. Therefore, we will focus the further analysis on the last two years: 01/11/2011–31/10/2012 and 01/11/2012–31/10/2013.

Fig. 3 presents the sentiment indexes for the two last years. The index fluctuates a lot with time: dynamic range of the word sentiment index is 5.5 (min/max), while dynamic range of the text sentiment index is 6. We used day-based plots as we found that week- and month-based plots are less informative. As one may notice, positive/negative bursts are short and thus smoothed easily. The extreme values of the indexes indeed coincide with the some important events in Russia and in the World:

(1) 31-12-*—New Year (+);
(2) 14-02-*—St.Valentine's day (+);
(3) 23-02-*—Man's day (+);
(4) 08-03-*—Woman's day (+);
(5) 09-05-*—Victory Day, World War 2 commemorative day (–);
(6) 07-07-2012—Krasnodar Krai floods in Russia[6] (–);
(7) 22-07-2012—A new unpopular law regulating non-profit organizations in Russia[7] (–);
(8) 16-09-2012—A mass protest against government in Russia[8] (–);
(9) 25-10-2012—Hurricane Sandy in US (–).

The list above merely reflects our interpretation of the sentiment index. While points (1)–(5) seem to be correct, extreme values of the index around points (6)–(9) can be due to a different cause. We did not have resources to validate our interpretations, as it would require manual check of a huge number of texts. For instance, to verify point (7) one would need to check topic of 17,490 texts issued on the 22/07/2012. However, it is clear that further interpretation of the results is desirable.

Fig. 4 depicts the emotional indexes for the two last years. The fraction of emotional texts e_t ranges between 3% and 5%. This quantity is relatively stable and has practically no huge bursts. The fraction of emotional words e_w varies in a similar way: it ranges from 0.15% until 0.35%. Dynamic range of both emotional indexes is about 2.2. One can observe a significant discrepancy of e_t and e_w during the last two years (Fig. 3) and during the whole period (Fig. 2). These differences are again due to the noisiness of the data during the period 2006–2010 (see Fig. 1 and 2). Some further observations are as follows:

- Emotional level of the social network varies periodically. The word and text emotional indexes (e_w and e_t) vary weekly (see Fig. 4 and 5). Fewer text are issued during weekends; fewer *emotional* texts are issued during weekends.
- Direction of emotions varies irregularly (see Fig. 3); extreme values of word and text sentiment indexes (s_w and s_t) coincide with some important events, such as New Year or a huge natural disaster.

---

[6] http://en.wikipedia.org/wiki/2012_Krasnodar_Krai_floods

[7] http://rusnewsjournal.com/menu/186/

[8] http://www.nytimes.com/2012/09/16/world/europe/anti-putin-protesters-march-in-moscow-russia.html?_r=0

It is clear that grouping results by sociodemographic factors can be of interest. This will let approach questions such as “If people from Moscow are in average more positive than people from Saint Petersburg?” or “If older people are more positive than youth?”, etc. However, in our pilot study we limit ourselves to measuring “global” indexes that average sentiment of all users.

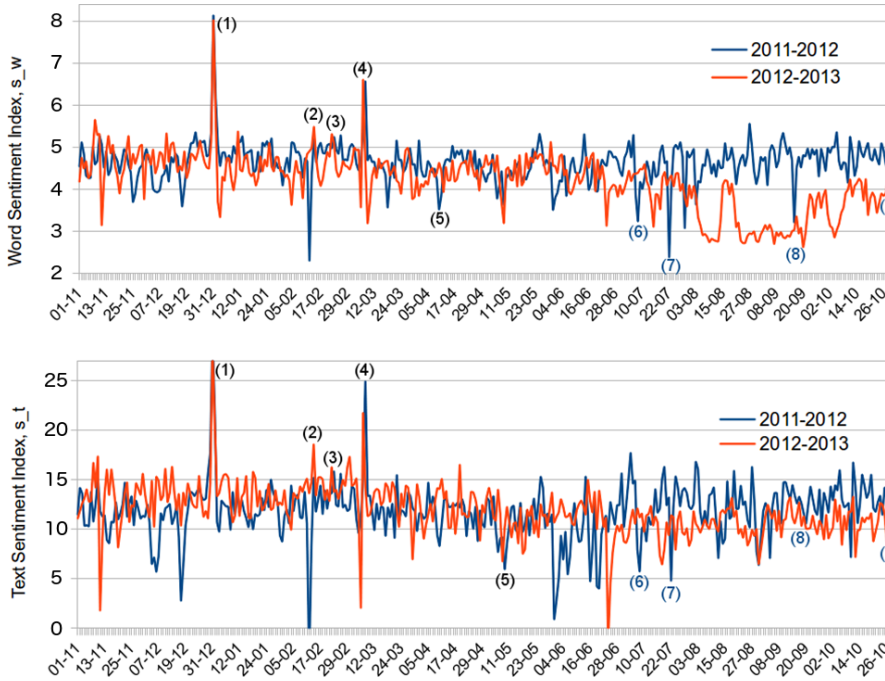


Fig. 3. Word (s_w) and text (s_t) sentiment indexes temporal evolution during two years: 11/2011—11/2012 and 11/2012—11/2013

4. Conclusion and Future Work

This paper presented some preliminary results on the social sentiment index of the Russian-speaking Facebook. First, we introduced four indexes that measure average emotional level in a social network. Two of them rely on sentiment words, while two other exploit text sentiment classification. We applied the proposed metrics on a corpus of Facebook posts. According to our results: (1) positive posts and comments predominate over negative ones; (2) maximum values of the index coincide with the national holidays, while minimum values coincide with national tragedies and commemoration days.

In further work, we plan to introduce a more sophisticated version of the index taking into account other factors, such as influence of a post and its author. We also would like to perform further analysis of the results to answer the questions such as: “Which time of the day is the most positive?” or “Are women use more positive terms than men?”.

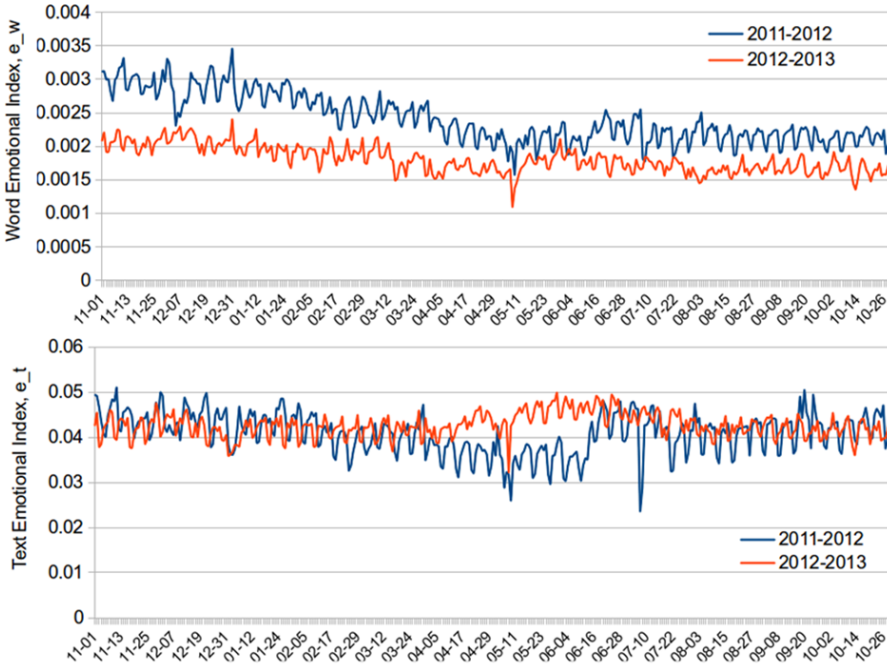


Fig. 4. Word (e_w) and text (e_t) emotional indexes temporal evolution during the two years: 11/2011—11/2012 and 11/2012—11/2013

Acknowledgements

This research was supported by Digital Society Laboratory LLC. We thank Sergei Objedkov and three anonymous reviewers for their helpful comments that significantly improved quality of this paper.

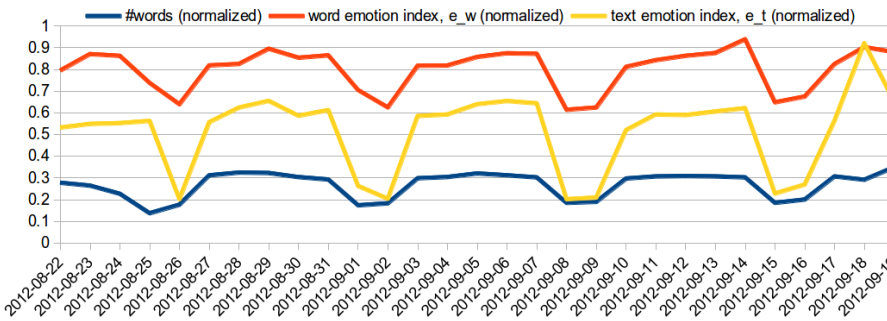


Fig. 5. Word (e_w) and text (e_t) emotional indexes, compared to the number of words. Each variable was projected to the range [0;1]

References

1. *Godbole, Namrata, Manja Srinivasaiah, and Steven Skiena.* "Large-Scale Sentiment Analysis for News and Blogs." ICWSM 7, 2007.
2. *Lui, Marco, and Timothy Baldwin.* "langid. py: An off-the-shelf language identification tool." Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics, 2012.
3. Facebook: Key Facts: http://newsroom.fb.com/Key-Facts
4. Internet World Stat: Internet and Facebook Usage in Europe (2012) http://www.internetworldstats.com/stats4.htm#europe
5. *Dodds, Peter Sheridan, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth.* "Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter." PloS one 6, no. 12 (2011): e26752.
6. *Pang, Bo, and Lillian Lee.* "Opinion mining and sentiment analysis." Foundations and trends in information retrieval 2. 1–2 (2008): 1–135.
7. *Liu, Bing.* "Sentiment analysis and subjectivity." Handbook of natural language processing 2 (2010): 568.
8. *Thelwall, Mike, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas.* "Sentiment strength detection in short informal text." Journal of the American Society for Information Science and Technology 61, no. 12 (2010): 2544–2558.
9. *Blinov P., Klekovkina M., Kotelnikov E., Pestov O.* Research of lexical approach and machine learning methods for sentiment analysis. In Proceedings of Dialog, Bekasovo, 2013
10. *Chetviorkin, Ilia, and Natalia Loukachevitch.* "Evaluating Sentiment Analysis Systems in Russian." ACL 2013 (2013): 12.
11. *Catanese, S., De Meo, P., Ferrara, E., Fiumara, G., & Provetti, A.* (2012). Extraction and analysis of Facebook friendship relations. In Computational Social Networks (pp. 291–324). Springer London.
12. *Mihalcea, Rada, and Hugo Liu.* "A Corpus-based Approach to Finding Happiness." AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. 2006.
13. *Kramer, Adam DI.* "An unobtrusive behavioral model of gross national happiness." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2010.
14. *Choi, Hyunyoung, and Hal Varian.* "Predicting the present with google trends." Economic Record 88. s1 (2012): 2–9.
15. *Kamvar, Sepandar D., and Jonathan Harris.* "We feel fine and searching the emotional web." Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011.
16. *Wang, N.,* et al. "Can well-being be measured using Facebook status updates? Validation of Facebook's Gross National Happiness Index." Social Indicators Research 115.1 (2014): 483–491.
17. *Mishne, Gilad, and Maarten de Rijke.* "Capturing Global Mood Levels using Blog Posts." AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. 2006.
18. *Balog, Krisztian, and Maarten de Rijke.* „Decomposing bloggers' moods." World Wide Web Conference. 2006.

УПРАВЛЕНИЕ ИНОЯЗЫЧНЫХ НЕОЛОГИЗМОВ — НАЗВАНИЙ ОБЪЕКТОВ КИНОИНДУСТРИИ

Пестова А. Р. (pestova2012@gmail.com)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Ключевые слова: заимствования, неологизмы, вариативность, именное управление

GOVERNMENT OF THE BORROWED NEOLOGISMS DENOTING OBJECTS OF FILM INDUSTRY

Pestova A. R. (pestova2012@gmail.com)

Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

The present paper deals with the government of borrowed neologisms, denoting objects of film industry: *трейлер* ‘trailer’, *тизер* ‘teaser’, *ремейк* ‘remake’, *сиквел* ‘sequel’, *приквел* ‘prequel’, *триквел* ‘trequel’, *квадриквел* ‘quadriquel’, *мидквел* ‘midquel’ and *интерквел* ‘interquel’. Dictionaries don’t give any information about syntactical features of these words. The study shows that government of these nouns is variational and the revealed constructions are synonymous and redundant. As language tends to eliminate redundancy, we tried to find the most popular variant for each word. The Statistics of Internet resources “Yandex.Novosti” (news segment), “Yandex.Blogi” (blogosphere) and corpus RuTenTen was analysed. All listed nouns tend to govern non-prepositional genitive. The used method can be applied to other borrowed neologisms, for example to nouns, referring to music scene (*ремикс* ‘remix’, *кавер* ‘cover version’, *(видео)клип* ‘video clip’ and *(видео)ролик* ‘video clip’). They prefer prepositional-case construction *на* + accusative.

Keywords: borrowed words, neologisms, variation, noun government

Управление — важная лингвистическая информация о слове. Тем не менее, в традиционных словарях русского языка оно представлено фрагментарно и непоследовательно, главным образом в немногочисленных иллюстративных

примерах, не отражающих вариативность (часто существующую в рамках нормы)[1].

Статья посвящена управлению слов, обозначающих объекты музыкальной и киноиндустрии: *трейлер, тизер, ремейк, сиквел, приквел, триквел, квадриквел, мидквел,* и *интерквел*. Актуальность этой работы обусловлена следующими факторами: 1) отсутствием информации об управлении этих слов в словарях; 2) вариативностью их управления, которая наблюдается в современном употреблении; 3) трудностями, которые испытывают носители русского языка при употреблении этих слов в словосочетании[2]; 4) шире — крайней малочисленностью научных работ, посвящённых управлению иноязычных неологизмов[3].

Рассмотрим тенденции синтаксического поведения перечисленных слов, разделив их на следующие подгруппы: ‘рекламные ролики фильма’ (*трейлер, тизер*), ‘новая серия фильма’ (*сиквел, приквел, триквел, квадриквел, мидквел*) и ‘новая версия фильма’ (*ремейк*).

1. ‘Рекламные ролики фильма’

1.1. Трейлер

Трейлер — это «рекламный ролик, составленный из наиболее зрелищных кадров нового кинофильма, нередко со специально доснятым материалом»[4]. Первое вхождение слова *трейлер* в этом значении в НКРЯ — 2003 г.[5], в сегменте «Яндекс.Новости» — 2000 г.[6]

Встречаются следующие варианты управления этого заимствования:

чего-л.:

(1) *Потом мне довелось увидеть* ***трейлер фильма****, и я поверил, что это комедия* («Московский комсомолец», 03.06.2013).

---

[1] Ю. А. Бельчиков и Г. Я. Солганик, рассматривая различные типы лексикографических изданий, указывают на то, что словари с информацией о синтаксических связях слов — это «наименее разработанный в русской лексикографии тип словарей» [Бельчиков, Солганик 1997: 41].

[2] Свидетельством этих трудностей являются, например, вопросы в «Справочную службу русского языка» Института русского языка им. В. В. Виноградова РАН: «Саундтрек фильма или из фильма?» (вопрос был задан 10 апреля 2009 г.); «Как правильно: Трейлер фильма или трейлер к фильму?» (вопрос был задан 6 февраля 2012 г.).

[3] См., например, [Валиахметова 2011], [Валиахметова 2012], [Шмелёва 2012].

[4] НСИС. Здесь и далее расшифровку аббревиатур см. в приложении «Словари и словарные материалы».

[5] ***Трейлеры*** *по 35–40 секунд, где еще в дело не вник, а уже конец* («Хулиган», 15.12.2003).

[6] *На сайте … появился смонтированный фанатами эпопеи* ***«трейлер»*** *ко второму эпизоду* (Кинопортал «Фильм.ру», 15.02.2000).

к чему-л.:

(2) *…на сайте YouTube … можно обнаружить как* ***трейлеры к фильму****, так и полноценные версии скандальной киноленты* («Российская газета», 26.12.2012).

на что-л.:

(3) *Несколько лет назад эксцентричный итальянец шокировал жюри Венецианской биеннале, представив на их суд собственную версию* ***трейлера на фильм*** *«Калигула»* («Vogue», 21.06.2013).

для чего-л.:

(4) *На следующей неделе выйдет первый* ***трейлер для фильма*** *«Тор: Темный мир»* (Сайт, посвященный кино, 2013).

по чему-л.:

(5) *Просмотрела* ***трейлер по фильму*** *этому — вот даже не знаю, качать или нет* (Блоги, 2011).

Рассмотренные конструкции дублетны: они встречаются в синонимичных контекстах и заменяют друг друга в пределах одного текста:

(6) *Дублированный* ***трейлер мультфильма*** *«Храбрая сердцем». Кинокомпания Дисней выпустила дублированный* ***трейлер к своему новому мультфильму*** *«Храбрая сердцем»…* (Сайт, посвященный кино, 2011).

Чтобы определить наиболее употребительный вариант управления, мы обращались к статистике следующих интернет-источников: сегментов «Яндекс.Новости» и «Яндекс.Блоги», а также корпуса RuTenTen[7]. И в новостных изданиях, и в блогах *трейлер* чаще всего управляет беспредложным родительным падежом. На втором месте по частотности — конструкция *к чему-л*. Примеры с управлением *на что-л.*, *для чего-л.* и *по чему-л.* встречаются редко, в основном в блогах и на форумах[8].

---

[7] Поиск всех словосочетаний в сегментах «Яндекс.Новости» и «Яндекс.Блоги» осуществлялся в декабре 2013 г., в корпусе RuTenTen — в апреле 2014 г.

[8] В сегментах Яндекса водились запросы вида запросы вида трейлер /+1 "фильма", трейлер /+1 "к фильму" и т. д. Поиск в сегменте «Яндекс.Новости» ограничен периодом 16.09.2013–16.12.2013, а в блогосфере — периодом 13.12.2013–16.12.2013, т. к. при поиске без такого ограничения выдаётся слишком много повторяющихся либо неадекватных запросу примеров.

Таблица 1. *Трейлер*: статистическая информация по данным интернет-источников «Яндекс.Новости», «Яндекс.Блоги» и RuTenTen

Интернет-источник	***фильма***	***к фильму***	***на фильм***	***для фильма***	***по фильму***
«Яндекс.Новости»	1119	416	1	0	0
«Яндекс.Блоги»	546	382	4	1	2
RuTenTen	1732	1266	15	99	6

1.2. Тизер

Этого неологизма в словарях русского языка нет. Тем не менее, в современных СМИ, блогах и устной речи он активно используется. Чаще всего *тизерами* называют видеоролики с самыми интересными и интригующими кадрами нового кинофильма. Первые вхождения в НКРЯ и в сегменте «Яндекс.Новости» датируются 2002 г.[9]

Слово *тизер* употребляется с четырьмя вариантами управления:

чего-л.:

(7) *А пока мы предлагаем вашему вниманию* ***тизеры короткометражек****, которые были сняты в Мурманской области* («Комсомольская правда в Мурманске», 06.02.2013).

к чему-л.:

(8) *Народный артист России Георгий Обухов сыграет роль профессора Брамса в* ***тизере к фильму*** *Сэнди Кролика* («Московский комсомолец», 24.09.2013).

на что-л.:

(9) *Вот* ***тизер на фильм*** *«Домовой» с Константином Хабенским и Владимиром Машковым* (Блоги, 2008).

для чего-л.:

(10) *Первый* ***тизер для фильма*** *«Джон Картер» от студии Disney, а также последовавший за ним трейлер не вдавались в детали того, как ветеран гражданской войны Джон Картер … попадает на Марс* (Сайт, посвященный кино, 2011).

---

[9] В НКРЯ: *В Голливуде такие штуки называются* ***тизерами****: это короткие рекламные ролики с самыми завлекательными моментами фильма* («Известия», 22.09.2002). В сегменте «Яндекс.Новости»: *…там же на сайте расположен* ***тизер****, за каждый клик по которому компания-спонсор переводит на программу по спасению алтайских лесов 6 центов* (Интернет-издание «Утро.ru», 18.10.2002).

Все эти варианты употребляются в синонимичных контекстах и иногда встречаются в пределах одного абзаца:

(11) *Первый **тизер для фильма** «Воскрешение Темного рыцаря». В сети появился первый **тизер фильма** «Воскрешение Темного рыцаря», который мы рекомендуем к просмотру нашим зрителям* (Сайт, посвященный кино, 2011).

Чаще всего, согласно статистике сегментов «Яндекс.Новости» и «Яндекс.Блоги», употребляется вариант с беспредложным родительным падежом. Корпус RuTenTen демонстрирует преобладание предложно-падежной конструкции *к чему-л.*, однако при вычитке примеров обнаруживается, что больше половины из них (76 из 124) представляют собой шаблонное предложение с сайта «В кинокресле», которое начинается словами «Смотреть трейлер к фильму…».

Реже всего встречаются предложно-падежные конструкции *на что-л.* и *для чего-л.*

Таблица 2. *Тизер*: статистическая информация по данным интернет-источников «Яндекс.Новости», «Яндекс.Блоги» и RuTenTen

Интернет-источник	***фильма***	***к фильму***	***на фильм***	***для фильма***
«Яндекс.Новости»	543	221	1	2
«Яндекс.Блоги»[10]	233	133	1	2
RuTenTen	98	48	1	0

Итак, семантическая подгруппа слов, обозначающих рекламные ролики фильма, используется с беспредложным родительным падежом и предложно-падежными конструкциями *к чему-л., на что-л., для чего-л.* и (только *трейлер*) *по чему-л.* При этом самым частотным для этих слов является вариант с беспредложным родительным падежом.

2. ‘Новая серия фильма’

2.1. Сиквел

Слова *сиквел, приквел, триквел, квадриквел, мидквел* и *интерквел* обозначают либо продолжение, либо предысторию какого-либо произведения. Чаще всего они обозначают новые серии фильма, поэтому мы рассматриваем их в этой семантической подгруппе.

[10] Учтена только статистика за период 13.11.2013–13.12.2013, т. к. при поиске без такого ограничения выдаётся слишком много повторяющихся либо неадекватных запросу примеров.

Сиквел — это «продолжение книги, романа; следующая серия многосерийного фильма, снятая, как правило, на волне успеха предыдущих серий»[11]. Первое вхождение этого неологизма в НКРЯ — 2000 г.[12], в сегменте «Яндекс.Новости» — 1993 г.[13]

Встречается пять вариантов управления существительного *сиквел*:

чего-л.:

(12) ***Сиквел фильма*** *Эльдара Рязанова два года удерживал титул самого кассового фильма российского проката…* («Комсомольская правда», 12.01.2010).

к чему-л.:

(13) *Своего рода* ***сиквелом к фильму*** *станет фильм «Сезон что надо»…* («Российская газета», 16.01.2013).

на что-л.:

(14) *Эта история — снятый режиссером Энди Фикманом* ***сиквел на фильм*** *1975 года «Побег на Ведьмину гору»* (Интернет-журнал, 2009).

для чего-л.:

(15) *«Бросок кобры 2» стал вполне достойным* ***сиквелом для*** *известного* ***боевика*** (Блоги, 2013).

по чему-л.:

(16) *Очередной* ***сиквел по фильму*** *Рязанова появится зимой следующего года* («Комсомольская правда», 05.04.2010).

Чаще всего встречается вариант управления беспредложным родительным падежом. На втором месте по частотности — предложно-падежная конструкция *к чему-л*. Остальные конструкции являются периферийными, см. табл. 3[14]:

---

[11] НСИС.

[12] *Карлофф позднее не один раз снимался в* ***сиквелах*** *— своего рода продолжениях — «Франкенштейна» и стал популярным актером именно в фильмах ужасов* (В. Быков, О. Деркач. Книга века).

[13] *Вовсе не обязательно, чтобы картины были прямыми «римейками» или «****сиквелами****», то есть продолжениями, как «Кладбище домашних животных II»* («Коммерсантъ», 05.05.1993).

[14] Поиск в сегменте «Яндекс.Новости» ограничен периодом 16.12.2010–16.12.2013, а в блогосфере — периодом 16.09.2013–16.12.2013.

Таблица 3. *Сиквел*: статистическая информация по данным интернет-источников «Яндекс.Новости», «Яндекс.Блоги» и RuTenTen

Сегмент	*фильма*	*к фильму*	*на фильм*	*для фильма*	*по фильму*
«Яндекс.Новости»	1426	398	0	0	2
«Яндекс.Блоги»	883	67	1	0	0
RuTenTen	239	48	1	0	1

2.2. Приквел

Приквел — это «произведение (литературное, кинематографическое, сценическое), повествующее о предыстории событий, уже известных читателю или зрителю по какому-л. роману, фильму и др.»[15].

В НКРЯ самое раннее вхождение этого слова датируется 2002 г.[16], в сегменте «Яндекс.Новости» — 1993 г.[17]

Приквел может управлять следующими конструкциями:

чего-л.:

(17) ***Приквел фильма*** *«Игры патриотов» выйдет в конце 2013 года* («РИА Новости», 23.08.2012).

к чему-л.

(18) *Первая игра из серии об убийце из будущего Риддике … стала своеобразным* ***приквелом к фильму*** *«Хроники Риддика»* (Новостной сайт, 2013).

для чего-л.:

(19) *Это издание выполняет роль своеобразного* ***приквела для фильма*** *«Top 2: Тёмный мир»* (Сайт, посвященный кино, 2013).

Наблюдается конкуренция двух вариантов управления: беспредложным родительным падежом и предложно-падежной конструкцией *к чему-л.* См. табл. 4:

Таблица 4. *Приквел*: статистическая информация по данным интернет-источников «Яндекс.Новости», «Яндекс.Блоги» и RuTenTen

Сегмент	*фильма*	*к фильму*	*для фильма*
«Яндекс.Новости»	317	225	1
«Яндекс.Блоги»	786	794	0
RuTenTen	44	24	0

[15] НСИС.

[16] *А для тех, кто полюбил «Мумий», —* ***приквел*** *«Царь скорпионов»* («Домовой», 04.04.2002).

[17] *…ох, уж эти американцы с их непреодолимой страстью к римейкам, сиквелам,* ***приквелам*** *и пр.* («Коммерсантъ», 20.11.1993).

2.3. Триквел, квадриквел, мидквел, интерквел

Кроме довольно распространённых слов *сиквел* и *приквел*, встречаются также более редкие *триквел* — «третье из серии последовательных произведений», *квадриквел*— «четвёртое произведение в серии» и *мидквел* — «произведение, развивающее сюжет предшествующих произведений на ту же тему» и *интерквел* — «художественное произведение, сюжетные события которого происходят между событиями ранее созданных произведений»[18]. Самые ранние примеры в НКРЯ со словами *триквел* и *квадриквел* — 2007 г.[19], вхождений со словами *мидквел* и *интерквел* в Корпусе нет вообще. В «Яндекс.Новостях» *триквел* и *квадриквел* встречаются впервые в 2002 г.[20], *мидквел* — в 2011 г.[21], *интерквел* — в 2012 г.[22]

Все эти неологизмы употребляются с беспредложным родительным падежом:

(20) *Том Селлек подтвердил слухи о подготовке* ***триквела фильма*** *«Трое мужчин и младенец», ставшего суперхитом в середине 1980-х* (Сайт, посвященный кино, 2010).

(21) *Кутерьма вокруг* ***квадриквела фильма*** *о похождениях секретного агента Джейсона Борна продолжается* (Блоги, 2010).

(22) *«300 спартанцев: Расцвет империи» —* ***мидквел фильма*** *«300 спартанцев» 2006 года* (Блоги, 2013).

(23) *Новый фильм начинается как* ***интерквел*** *самого первого «****Повелителя кукол****»* (Блоги, 2012).

Триквел и *мидквел* встречаются, кроме того, с вариантом *к чему-л.*:

(24) *Вспомните «Крик 3», в котором неудачник-брат Сидни снимал* ***триквел к фильму*** *«Удар ножом»* (Блоги, 2011).

---

[18] «Википедия».

[19] *Ему и предстоит определить, какой фильм станет самым кассовым* ***триквелом*** *этого года* («РБК Daily», 03.04. 2007).
Актеры играют с потухшими глазами, так, что сразу видно — они потеряли интерес к ***квадриквелу*** *еще до того, как прозвучала команда «Мотор!»* («РБК Daily», 05.03.2007).

[20] ***Триквел*** *блокбастера «Матрица. Революции» … выйдет в ноябре 2003 года* («Коммерсантъ-Online», 09.08.2002).
Зритель валялся в саспенсе, как свинья в навозе: было понятно, что последует сиквел, приквел и ***квадриквел*** («Грани.ру», 16.05.2002).

[21] *Если бы киноманы оценили все четыре части, можно было бы покорпеть и над* ***мидквелом*** *«Мой левак»…* (Интернет-издание «Багнет», 11.02.2011).

[22] *Спин-офф, в зависимости от того, в какое время относительно оригинала происходят его события, может быть как сиквелом, так и приквелом, и мидквелом, и* ***интерквелом****…* («Псковская правда», 09.04.2012).

(25) *Возвращение на Алу — **мидквел к роману** Изгнанник вечности…* (Развлекательный портал, 2013).

Значительно преобладают у этих слов, тем не менее, вариант с беспредложным родительный падежом. См. статистику сегментов «Яндекс.Новости» и «Яндекс.Блоги» для слова *триквел*[23]*:*

Таблица 5. *Триквел* и *мидквел*: статистическая информация по данным интернет-источников «Яндекс. Новости», «Яндекс.Блоги» и RuTenTen

Интернет-источник	***триквел фильма***	***триквел к фильму***	***мидквел фильма***	***мидквел к фильму***
«Яндекс.Новости»	30	1	2	0
«Яндекс.Блоги»	52	8	3	1
RuTenTen	6	2	0	1

Итак, в рассмотренной семантической подгруппе ‘новая серия фильма’ слова *квадриквел*, и *интерквел* употребляются только с беспредложным родительным падежом. *Сиквел*, *триквел* и *мидквел* встречаются и с другими конструкциями, но предпочитают этот же вариант. Что касается слова *приквел*, то его управление колеблется между двумя вариантами: беспредложным родительным падежом и предложно-падежной конструкцией *к чему-л*. Однако, учитывая поведение других слов этой семантической подгруппы, можно предположить, что перспективным является именно первый вариант.

3. ‘Новая версия фильма’

Ремейк— это «новая, исправленная или восстановленная версия старого фильма, спектакля, музыкальной записи и др.»[24]. Первое вхождение в НКРЯ — 1996 г. (в варианте написания *римейк*)[25] и 1997 г. (в варианте написания *ремейк*)[26]; в сегменте «Яндекс.Новости» — 1993 г. (оба орфографических варианта)[27].

---

[23] См. выше.

[24] НСИС.

[25] *На ОРТ шел фильм «Старые песни о главном» — пародийный **римейк** сразу всех советских **фильмов** про колхозную деревню — «Кубанских казаков», «Свадьбы с приданым», «Богатой невесты» и прочих* («Коммерсантъ-Daily», 20.01.1996).

[26] *На большом экране хиты малобюджетного независимого американского кино 1996 года: … **ремейк «Телохранителя»** Куросавы «Герой-одиночка» ….* («Столица», 01.07.1997).

[27] *…футбольный матч на Красной площади — своего рода **римейк** матча 1936 г.…* («Коммерсантъ», 02.03.1993); *…автор **ремейка** — знаменитый петербургский театральный художник и коллекционер русского модерна Вячеслав Окунев* («Коммерсантъ», 06.03.1993).

В иллюстративных примерах, приведённых в словарях, зафиксировано два варианта управления: *на что-л.* и *чего-л.*: *Кульминацией вечера стало появление «DeadУшек», которые в рамках своей традиционной программы исполнили и* ***ремейки на песни*** *Бориса Гребенщикова, причем на сей раз вступал вокал Виктора Сологуба*. ПТ, 15.02.99–21.02.99[28]; ***Ремейк этого фильма*** *уже вышел на экраны*[29].

В современном употреблении встречается пять вариантов:

чего-л.:

(26) *Фильму было уже несколько лет, моей посетительнице едва ли 30, и мне стоило немалого труда убедить ее, что «Сабрина» — это вообще* ***римейк картины*** *1954 года* («Домовой», 04.02.2002).

на что-л.:

(27) *Одно дело, если делается* ***римейк на*** *какое-то* ***произведение*** *или делается новая аранжировка, и совсем другое, когда берутся фрагменты и на эту тему делается новое произведение* («Известия», 18.01.2002).

к чему-л.:

(28) *Оказалось, права на* ***ремейк к фильму*** *«Москва слезам не верит» режиссер давно продал* (Новостной сайт, 2011).

по чему-л.:

(29) *Еще более знаменательным обстоятельством стало для режиссера намерение одной из американских кинокомпаний создать* ***римейк по фильму*** *«Пленный», перенеся события из Чечни в Ирак* (Сайт телеканала «Культура», 05.09.2008).

для чего-л.:

(30) *Голливудский актер Киану Ривз напрочь отказался сниматься в* ***ремейке для фильма*** *«На гребне волны»* (Новостной сайт, 2013).

Самой частотной является конструкция с беспредложным родительным падежом. Так, в НКРЯ в 58 из 61 вхождения со словом *ремейк / римейк* в роли управляющего слова используется этот вариант[30]; в остальных трёх примерах оно управляет конструкцией *на что-л*. Примеры с конструкциями *к чему-л*., *по чему-л*. и *для чего-л.* в НКРЯ не встречаются. Аналогичный результат даёт поиск по сегментам «Яндекс.Новости» и «Яндекс.Блоги», см. нижеприведённые таблицы[31].

---

[28] АЛ.

[29] 1000 НИС.

[30] Поиск осуществлялся 27.01.2012.

[31] Поиск осуществлялся в декабре 2013 г. Поиск в сегменте «Яндекс.Новости» ограничен периодом 16.12.2012–16.12.2013, а в блогосфере — периодом 16.11.2013–16.12.2013, т. к. при поиске без такого ограничения выдаётся слишком много повторяющихся либо неадекватных запросу примеров.

Сегмент	*ремейк/ римейк фильма*	*ремейк/ римейк на фильм*	*ремейк/ римейк к фильму*	*ремейк/ римейк по фильму*	*ремейк/ римейк для фильма*
«Яндекс. Новости»	376 / 161	16 / 2	3 / 1	0 / 1	1 / 0
«Яндекс. Блоги»	216 / 91	22 / 6	1 / 0	0 / 0	3 / 0

RuTenTen

ремейк/ римейк фильма	*ремейк/ римейк на фильм*	*ремейк/ римейк к фильму*	*ремейк/ римейк по фильму*	*ремейк/ римейк для фильма*
24 / 23	21 / 20	5	1 / 2	0

Выводы

Изучив управление иноязычных неологизмов, обозначающих объекты киноиндустрии, можно отметить следующее. Варианты управления этих слов дублетны: словосочетания с разными конструкциями используются в синонимичных контекстах и иногда взаимозаменяются в пределах одного текста. Таким образом, вариативность управления рассмотренных неологизмов избыточна. Как известно, язык стремится избавиться от избыточности, а потому у этих слов наблюдается явное предпочтение одного варианта из нескольких. Все эти слова предпочитают беспредложное управление родительным падежом, вероятно, под действием семантической аналогии: ср. *трейлер, тизер фильма — реклама фильма*, *сиквел, приквел, триквел, квадриквел, мидквел, интерквел фильма — серия фильма*, *ремейк фильма — переделка, новая версия фильма*.

Подчёркивая неполноту синтаксической информации в словарях, отметим, что представленный в статье метод может быть применён и для выявления частотных вариантов управления других неологизмов. В частности, аналогичным способом были определены предпочтительные конструкции, которыми управляют иноязычные слова — названия объектов музыкальной индустрии: *ремикс, кавер, (видео)клип* и *(видео)ролик*.

Неологизмы со значением ‘новая версия песни’ (*ремикс* и *кавер*) чаще всего употребляются с вариантом *на что-л.*: *ремикс на песню, кавер на композицию*. Интересно, что здесь нет ожидаемого действия семантической аналогии со словами *версия* и *переделка*, как в случае со словом *ремейк*. Зафиксированы также примеры с управлением беспредложным родительным падежом и предложно-падежными конструкциями *на что-л., к чему-л., для чего-л.* и (только *ремикс*) *по чему-л.*

Эту же конструкцию (*на что-л.*) предпочитают и неологизмы, обозначающие музыкальные видеоролики: *(видео)ролик на песню, (видео)клип на композицию*. Они встречаются, кроме того, со следующими вариантами: беспредложным родительным падежом и предложно-падежными конструкциями *к чему-л., с чем-л., под что-л., для чего-л.* и *по чему-л.*

Литература

1. *Бельчиков Ю. А., Солганик Г. Я.* О лексикографических изданиях адресной направленности // Облик слова. Сборник статей памяти Дмитрия Николаевича Шмелёва. — М.: Институт русского языка РАН, 1997. — С. 41–47.
2. *Валиахметова А. Р.* Вариативность управления иноязычных существительных, называющих объекты Интернета // Вестник РГГУ. Серия «Филологические науки. Языкознание». Т. 13. — М.: Российский государственный гуманитарный университет, 2011. — С. 29–42.
3. *Валиахметова А. Р.* Управление иноязычных существительных — названий лиц по профессии в современном русском языке // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (30 мая — 3 июня 2012 г., Бекасово). Вып. 11 (18). Т. 1. — М.: РГГУ, 2012. — 761 с. — С. 638–651.
4. *Шмелёва Е. Я.* Делаем шопинг: о сочетаемости и семантике новых заимствований // Вопросы культуры речи. Вып. XI / Отв. ред. А. Д. Шмелёв. — М.: Языки славянской культуры, 2012. — С. 273–283.

Словари и словарные материалы

1. *1000 НИС* — Крысин Л. П. 1000 новых иностранных слов. — М.: АСТ-ПРЕСС КНИГА, 2009. — 320 с.
2. *АЛ* — Толковый словарь русского языка начала XXI века. Актуальная лексика / Под ред. Г. Н. Скляревской. — М.: Эксмо, 2007. — 1136 с.
3. *БТС* — Большой толковый словарь русского языка / Под ред. С. А. Кузнецова. Электронный ресурс. Режим доступа: http://www.gramota.ru/slovari/info/bts/, свободный. Загл. с экрана. Данные соответствуют 14.02.2014. Словарь опубликован в авторской редакции 2009 года.
4. *Википедия* — Википедия. Свободная энциклопедия. — Электронный ресурс. Режим доступа: http://ru.wikipedia.org, свободный. Загл. с экрана. Данные соответствуют 30.01.2014.
5. *НСИС* — Захаренко Е. Н., Комарова Л. Н., Нечаева И. В. Новый словарь иностранных слов. — М.: Азбуковник, 2008. 1040 с.

References

1. *Bel'chikov Ju. A., Solganik G. Ja.* Specific lexicographical editions [O leksikograficheskih izdanijah adresnoj napravlennosti], The face of word [Oblik slova], Institut russkogo jazyka RAN, Moscow, pp. 41–47.
2. *Shmelëva E. Ja.* (2012), Делаем шопинг 'Doing shopping': Combinability and semantics of new borrowed words [Delaem shoping: o sochetaemosti I semantike novyh zaimstvovanij], Speech culture issues [Voprosy cul'tury rechi], Issue 11, pp. 273–283.
3. *Valiahmetova A. R.* (2011), The Borrowed nouns denoting Internet objects: variation of the government [Variativnost' upravlenija inojazychnyh sushchestvitel'nyh, nazyvajushchih ob"ekty Interneta], RGGU Bulletin № 11 (73)/11 [Vestnik RGGU], Moscow, pp. 29–42.
4. *Valiahmetova A. R.* (2012), Government of the borrowed nouns denoting professions in the modern Russian language [Upravlenije inojazychnyh sushchestvitel'nyh — nazvanij lits po professii v sovremennom russkom jazyke], Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2012) [Komp'juternaja lingvistika i intellektual'nyje tehnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferentsii "Dialog"], Bekasovo, pp. 638–651.

ПРАГМАТИКА ЗАЧЁРКИВАНИЯ: НОРМЫ КОММУНИКАЦИИ И ТЕОРИЯ ОПТИМАЛЬНОСТИ

Пиперски А. Ч. (apiperski@gmail.com),
Сомин А. А. (somin@tut.by)

Российский государственный гуманитарный университет / ШАГИ РАНХиГС, Москва, Россия

В статье предлагается анализ преднамеренного зачёркивания в письменной речи с точки зрения различных лингвистических теорий: теории импликатуры Г. П. Грайса, теории вежливости П. Браун и С. Левинсона, а также теории оптимальности. Доказывается, что зачёркивание позволяет пролить свет на более общие механизмы коммуникации, в том числе на процесс выбора одной точки зрения из нескольких возможных. Кроме того, анализируются графические и вербальные аналоги зачёркивания в блогах и художественной литературе.

Ключевые слова: зачёркивание, язык Интернета, постулаты Грайса, теория вежливости, теория оптимальности, прагматика

PRAGMATICS OF STRIKETHROUGH: NORMS OF COMMUNICATION AND OPTIMALITY THEORY

Piperski A. Ch. (apiperski@gmail.com),
Somin A. A. (somin@tut.by)

Russian State University for the Humanities, Moscow, Russia

The paper presents a description of intentional strikethrough on the Web using a combination of theories from pragmatics and phonology, namely the theory of implicature by Grice (1975), the politeness theory by Brown and Levinson (1978, 1987), and Optimality Theory. We argue that the study of this phenomenon can shed light on some more general aspects of communication theory, such as the mechanism of choosing one viewpoint among many options. We also describe graphical and verbal substitutes for strikethrough in blogs and literary works.

Key words: strikethrough, language on the Web, Gricean maxims, politeness theory, Optimality Theory, pragmatics

1. Введение: самоисправление и его типы

Самоисправление является чрезвычайно распространённым языковым приёмом. Можно выделить два параметра, характеризующих самоисправление:

1) способ самоисправления: непреднамеренное / преднамеренное;

2) тип коммуникации: устный / письменный

В последние десятилетия всё большее внимание лингвистов привлекает непреднамеренное самоисправление в устной коммуникации, поскольку оно позволяет пролить свет на процесс порождения речи (ср. [Кибрик, Подлесская 2009] для русского языка; там же содержится библиография для других языков). Непреднамеренное зачёркивание в письменной коммуникации (т. е. обычное зачёркивание с целью исправления ошибок) не входит в сферу интереса лингвистов; о графологических работах, посвящённых этому вопросу, нам неизвестно.

Преднамеренное самоисправление возникает не в процессе естественного порождения устной или письменной речи, а как специальный приём языковой игры, имитирующий самоисправление: говорящий / пишущий заранее продумывает, что он скажет / напишет и на что тут же исправится. В риторике этот приём известен под разными названиями: эпанортосис, коррекция, эксполиция и т. п. (см. [Dupriez 1991: 165]).

Преднамеренное устное самоисправление встречается не слишком часто, однако примеры этого явления всё же существуют. Так, широкую известность получила фраза телеведущего Владимира Познера, который в своей программе сымитировал оговорку, якобы перепутав слова *дура* и *Дума*:

(1) *Вот способ мышления, который продемонстрировала Государственная дура — ой, простите, оговорился — Государственная Дума, ну, поражает…* (В. Познер, программа «Познер», 23.12.2012)

Однако естественно изобразить оговорку типа (1) довольно сложно: для этого нужно определённое актёрское мастерство, и именно поэтому преднамеренное устное самоисправление не является частотным.

Намного более распространён такой приём, как преднамеренное письменное самоисправление:

(2) *А еще сегодня под утро мне приснился наш ~~паноптикум~~ НИИ, в котором я якобы снова работаю* (из блога, 2009)

Обычно такие самоисправления имеют вид зачёркнутого текста, и поэтому далее мы будем называть этот приём просто словом «зачёркивание», по умолчанию подразумевая его преднамеренность.[1]

---

[1] В работе [Гусейнов 2008] для обозначения этого явления вводится термин «литуратив», однако мы не используем его, поскольку не считаем, что применение латинского термина привносит ясность в анализ.

2. Семантическая классификация зачёркиваний

Приём зачёркивания получил широкое распространение в русскоязычных блогах[2] в начале и середине 2000-х годов. Он не исчез и до сих пор, хотя и стал более редким, что связано с перемещением большей части публичной онлайн-коммуникации из блогосферы в социальные сети, которые технически не поддерживают зачёркивание (подробнее о техническом обеспечении приёма зачёркивания на разных платформах см. в [Пиперски, Сомин 2013: 615–617]).

Приём зачёркивания уже привлекал внимание исследователей. Зачёркиваниям в литературе были посвящены два семинара («Зачёркнутое слово в перспективе художественного высказывания» и «Зачёркивание как акт коммуникации»), прошедшие в 2012 и 2014 годах в Санкт-Петербурге в ИРЛИ РАН (Пушкинском доме): в большей части докладов рассматривались зачёркнутые фрагменты в черновиках художественных произведений, однако в некоторых докладах обсуждалось и зачёркивание как художественный приём (см. также раздел 8 настоящей статьи).

Зачёркиваниям в Интернете посвящены работы [Гусейнов 2008] и [Занегина 2009]: в них обсуждается классификация зачёркиваний с точки зрения семантики, т. е. какого рода высказывания зачёркиваются. В работе [Савицкая 2009] предпринимается попытка проанализировать зачёркивания с точки зрения прагматики, однако прагматическое описание этого явления оказывается чересчур общим, а анализ конкретных примеров — напротив, слишком частным и близким к семантическому.

Проблема семантического описания зачёркиваний заключается в том, что даже если считать, что описываются только наиболее частые типы, то при попытке расклассифицировать по семантическим основаниям большой корпус примеров всё равно всегда можно найти ряды схожих между собой высказываний, не входящих ни в один выделенный исследователем класс. С этой точки зрения нельзя не согласиться с заключением М. А. Кронгауза: «Лингвисты пытаются классифицировать типы зачёркивания, но это задача крайне неблагодарная, потому что зачеркнуть можно всё что угодно» [Кронгауз 2013: 208].

3. Зачёркивание: семантика или прагматика?

Правда, вывод М. А. Кронгауза может быть обобщён и шире: не только в случае зачёркивания, но и вообще **в любой ситуации** можно сделать всё что угодно. Но это не повод предаваться гносеологическому пессимизму и вовсе отказываться от описания человеческого поведения: хотя в любой ситуации можно сделать всё что угодно, существуют некоторые стандартные модели, определяющие поведение человека. Точно так же, как преподаватель во время лекции обычно не дарит студентам цветы, не исполняет матерные частушки

---

[2] Употребление зачёркивания в интернет-коммуникации на других языках требует дополнительного исследования.

и не красит лаком ногти, участник письменной коммуникации не зачёркивает что угодно и где угодно: он следует принятым нормам коммуникации и принимает решение о том, что зачеркнуть, исходя из своего представления о них.

Однако следует согласиться с М. А. Кронгаузом в том, что семантические классификации зачёркиваний — это списки интересных фактов, не ведущие к обобщению. Можно провести аналогию с геометрией: если заниматься изучением треугольников, можно перечислять их интересные свойства (теорема Пифагора, второй признак равенства треугольников, теорема Фалеса, формула Герона и т. п.), но с научной точки зрения важнее не перечни интересных свойств, а то, как эти свойства выводятся из ограниченного набора базовых аксиом[3]. Точно так же в нашей работе будут исследоваться не интересные, но частные значения зачёркивания, а то, как они выводятся из базовых норм коммуникации, касающихся не только зачёркивания, но и других аспектов речевой деятельности.

4. Нормы коммуникации

Мы выделяем два основных принципа, лежащих в основе человеческой коммуникации:

1) Принцип кооперации;

2) Принцип самоутверждения.

Принцип кооперации был сформулирован Г. П. Грайсом [Grice 1975, Грайс 1985]. Под этим термином понимается стремление коммуникантов способствовать успешному развитию диалога: «твой коммуникативный вклад на данном шаге диалога должен быть таким, какого требует совместно принятая цель (направление) этого диалога» [Grice 1975: 45, Грайс 1985: 222]. Принцип кооперации распадается на несколько более частных норм, или постулатов коммуникации: «Твоё высказывание должно содержать не меньше информации, чем требуется», «Твоё высказывание не должно содержать больше информации, чем требуется», «Не говори того, что ты считаешь ложным» и т. д. [Grice 1975: 45–46, Грайс 1985: 222–223].

Суть Принципа самоутверждения заключается в том, что любой участник диалога стремится не только к успешному достижению совместных целей диалога, но и к формированию своего положительного образа в глазах собеседника. Ближе всего к формализации этого принципа в прагматике стоит теория вежливости П. Браун и С. Левинсона [Brown & Levinson 1978, Brown & Levinson 1987]. В этой теории выделяются два аспекта положительного образа[4]:

---

[3] При этом аксиомы внешне выглядят гораздо менее увлекательно, чем выводимые из них теоремы: утверждение «Каковы бы ни были две точки *A* и *B*, существует прямая *a*, которой принадлежат эти точки» кажется намного более банальным, чем утверждение «Квадрат гипотенузы равен сумме квадратов катетов». Но без первого не было бы второго.

[4] Само понятие лица было введено ещё И. Гофманом [Goffman 1955], но именно П. Браун и С. Левинсон предложили разделять это понятие на две составные части.

1) «позитивное лицо» (*positive face*): положительный образ человека как добропорядочного члена общества;
2) «негативное лицо» (*negative face*): положительный образ человека как индивидуума, свободного от условностей.

В современной культуре (во всяком случае, российской и западной) значимы оба этих компонента образа[5]: для того, чтобы производить положительное впечатление, человек должен в определённых пропорциях сочетать в себе конформность (позитивное лицо) и независимость (негативное лицо): идеальный человек, обладающий максимально позитивным лицом и не обладающий негативным, встречается разве что в плохой литературе, но коммуницировать с ним едва ли было бы приятно.

Важным, если не важнейшим средством поддержания лица является язык. Однако взаимодействие языка с типами лица неоднозначно. Например, матерная брань может способствовать ухудшению позитивного лица, если подросток употребляет её на школьном уроке, а может, напротив, приводить к улучшению негативного лица, если дело происходит на школьном дворе в кругу сверстников.

Любое высказывание сложным образом соотносится с Принципом кооперации Грайса и с различными аспектами Принципа самоутверждения. Невозможно соблюсти все постулаты Грайса одновременно [Horn 2004: 8], а при этом ещё и улучшить и позитивное, и негативное лицо.

5. Теория оптимальности и нормы коммуникации

Ситуация, при которой имеется ряд взаимодействующих между собой запретов, некоторые из которых приходится нарушать, в лингвистике лучше всего описывается при помощи теории оптимальности [Prince & Smolensky 1993, Kager 1999]. Эта теория в первую очередь применяется к фонологии, а суть её сводится к тому, что для определения поверхностного представления той или иной фонологической единицы порождается некоторый набор кандидатов, который затем проверяется на соблюдение ранжированных ограничений: сперва выбывают все кандидаты, нарушившие самое важное ограничение, затем — все кандидаты, нарушившие второе по значимости ограничение, и так далее, пока не остаётся единственный кандидат, который и становится победителем (это может произойти ещё до того, как в дело вступили все ограничения).

Отметим два важнейших свойства теории оптимальности:

1) **победитель оптимален, но не идеален**: он может нарушать некоторые ограничения, которые, однако, оказываются малозначимыми, потому что до них не дошла очередь при проверке;
2) **ограничения могут ранжироваться по-разному** в разных грамматиках; при разном ранжировании будут побеждать разные кандидаты

[5] К другим культурам противопоставление типов лица применимо хуже; ср. обзор в [Вахтин, Головко 2004: 242–247].

(например, есть языки, в которых запрет на звонкий шумный согласный в конце слова ранжируется высоко, и в таких языках выигрывают кандидаты с глухим конечным согласным — иначе говоря, происходит оглушение конечных согласных; в других же языках это ограничение ранжируется низко, и победу одерживают кандидаты со звонким конечным согласным — иначе говоря, оглушения не происходит).

Аналогичным образом можно представить и процесс порождения высказывания с точки зрения прагматики: сперва генерируется набор высказываний, которые говорящий потенциально может произнести, а затем они проверяются на ненарушение норм коммуникации, образующих Принцип кооперации и Принцип самоутверждения («не говори неправду», «не говори слишком много», «не говори слишком мало», «не ухудшай своё позитивное лицо», «не ухудшай своё негативное лицо», «не ухудшай лицо собеседника» и т. д.). При этом ограничения могут ранжироваться по-разному у разных говорящих (например, у более многословных людей выше ранжируется запрет «не говори слишком мало», а у более лаконичных — запрет «не говори слишком много») и в разных коммуникативных ситуациях (например, в разговоре с начальником особенно важно не ухудшать своё позитивное лицо, а в разговоре с врачом — не говорить неправду о своём состоянии).

6. Порождение и интерпретация зачеркивания в теории оптимальности

Теория оптимальности применима и к ситуации с зачёркиванием. Однако в данном случае выбирается не один победитель, а два — «золотой» и «серебряный» медалист. «Золотой медалист» — это кандидат, который побеждает в соревновании по обычным правилам, однако после этого оценка кандидатов продолжается до тех пор, пока не выявляется кандидат, занявший второе место: именно он и отправляется под зачёркивание и ставится левее «золотого медалиста». В соревновании всегда участвует и нулевой кандидат (Ø = не говорить ничего); постулирование нуля позволяет описать зачёркивание без замены, например:

(3) *Босняки — это типа мусульман? По виду не скажешь — обычные люди* ~~*(без рогов и хвостов)*~~ (из блога, 2006)

Зачёркивание без замены — это победа нуля над остальными кандидатами, один из которых занимает второе место и зачёркивается (подробнее о зачёркиваниях с заменой и без см. [Пиперски, Сомин 2013]).

Теперь рассмотрим пример (4):

(4) *Как и любая порядочная* ~~*садистка*~~ *мать, я пользовалась каждым удобным случаем ознакомить ребенка с разнообразием животного мира* (из блога, 2008)

Здесь в конкуренцию вступают по крайней мере три точки зрения: *садистка*, *мать* и Ø. Возможно, автор рассматривал и других кандидатов, но об их существовании мы не можем узнать при анализе. Кандидат *садистка* нарушает ограничение «не ухудшай позитивное лицо» (нехорошо называть садистом самого себя; кроме того, называть садистами тех, кто знакомит детей с разнообразием животного мира, противоречит принятым в обществе представлениям), кандидат Ø нарушает запрет «не нарушай грамматическую правильность предложения», а кандидат *мать* не нарушает ничего. Судя по получившемуся тексту, оценивание кандидатов происходило следующим образом:

		*НарушГраммПрав	*УхудшПозЛиц
☞	*мать*		
❀	*садистка*		*!
	Ø	*!	

(в этой таблице * означает нарушение ограничения, ! — принятие решения об отсеивании кандидата; ☞ обозначает «золотого медалиста», а ❀ — «серебряного», попадающего под зачёркивание; более важные запреты стоят левее, чем менее важные, поэтому таблицу следует просматривать по столбцам слева направо; серая заливка обозначает ячейки, которые просматривать не нужно, потому что соответствующий кандидат уже исключен).

Рассмотрим ещё один пример:

(5) ~~*После долгой борьбы между совестью и жадностью*~~ *В результате научной дискуссии был решено вычеркнуть мыло «Хозяйственное» из исследования* (из блога, 2008)[6]

В этом примере проявляется взаимодействие ограничений «не говори неправду», «не ухудшай позитивное лицо» и «не говори слишком мало». Он анализируется так:

		*ГовориМало	*УхудшПозЛиц	*ГовориНеправду
❀	*долгая борьба*		*!	
☞	*научная дискуссия*			*
	Ø	*!		

При этом важно, что ограничения могли бы ранжироваться по-другому: если бы *УхудшПозЛиц и *ГовориНеправду шли в обратном порядке, зачёркнута была бы *научная дискуссия*, а не зачёркнута — *долгая борьба*. Таким образом, предлагаемая модель позволяет предсказать вариативность при зачёркивании там, где она есть, и не предсказывает её там, где она интуитивно

[6] Речь идёт о сравнении разных сортов мыла по заказу производителя одного из них. Опыты показали, что простое хозяйственное мыло превосходит все остальные сорта, в том числе и продукцию компании, заказавшей исследование.

кажется невозможной (напр., в примере (4) кандидат *мать* побеждает остальных кандидатов при любом ранжировании ограничений).

Типичный случай отсутствия вариативности — зачёркивание устойчивых выражений и фразеологизмов. Они появляются в тексте для повышения позитивного лица говорящего — он как бы сообщает адресату: «Мы с тобой одной крови, потому что я использую знакомые тебе фразеологизмы». Но при проверке запретов описание реальной ситуации никогда не проигрывает устойчивым выражениям, а при проверке на нарушение постулата истинности всегда выигрывает у них:

(6) *Вот сломанная детская игрушка, <...>, окаменевшие остатки пищи на столе, покрытым толстым-толстым слоем ~~шоколада~~*[7] *пыли…*
(из блога, 2005)

		*ГовориНеправду
❀	*шоколада*	*!
☞	*пыли*	

Разумеется, речь здесь не идёт об осознанном последовательном переборе кандидатов с оценкой их приемлемости — точно так же, как в фонологии говорящий не проводит осознанный оптималистский анализ при произнесении каждого слова. Напротив, примечательно, что механизмы зачёркивания реализуются неосознанно, хотя само зачёркивание и является намеренной языковой игрой.

Важно, что адресат может проанализировать зачёркивание не так, как его задумывал пишущий, поскольку он имеет дело со своего рода чёрным ящиком, из которого ему нужно вывести перечень наиболее значимых для говорящего запретов и их ранжирование. **Потенциальная множественность интерпретаций** — чрезвычайно характерное свойство зачёркивания как приёма языковой игры. Например, пример (2) … *~~паноптикум~~ НИИ* … можно трактовать так:

а) кандидат *паноптикум* появляется для соблюдения постулата полноты (для того, чтобы не только дать официальное обозначение места работы автора, но и выразить его отношение к нему) и зачеркивается, поскольку автор высоко ранжирует запрет на порчу позитивного лица;

б) кандидат *паноптикум* появляется для улучшения негативного лица (автор показывает, что может «с высоты птичьего полёта» посмотреть на место своей работы) и зачеркивается, поскольку автор высоко ранжирует постулат истинности (на самом деле это не *паноптикум*, а нормальный *НИИ*).

Как хорошее стихотворение не имеет единственно правильного понимания, так и удачное зачёркивание может быть проинтерпретировано поразному — и это не недостаток интерпретации, а сущность самого этого приёма языковой игры.

---

[7] Фраза из рекламы шоколадных батончиков Mars.

7. Проблема генерации кандидатов

Один из самых проблематичных вопросов теории оптимальности касается механизма генерации кандидатов. В стандартной версии теории предполагается, что множество кандидатов бесконечно, но при этом имеются механизмы, которые позволяют отсеять лишние — примерно так же, как при решении математического уравнения с одним корнем мы можем выбрать единственное решение из всего множества действительных чисел [Kager 1999: 25–27]. Однако применимость такой модели к прагматике вызывает большие сомнения. Кажется более разумным предполагать, что для каждого высказывания мы генерируем конечный набор кандидатов разной степени удачности, а затем проверяем их на нарушение ограничений.

Возникает вопрос о том, каким же образом происходит генерация кандидатов. Можно предположить, что говорящий генерирует кандидатов, подсознательно перебирая нормы коммуникации и стремясь соблюсти их. Если кандидатов сгенерировано более одного, происходит их оценка по описанным в разделе 6 принципам. Другими словами, мы сперва генерируем кандидатов, соблюдающих те или иные нормы коммуникации (например, находим кандидата, который позволяет соблюсти постулат истинности, другого кандидата, который удовлетворяет постулату количества, третьего кандидата, улучшающего наше позитивное лицо и т. п.), а затем отбрасываем те из них, которые нарушают наиболее важные запреты. Такая модель позволяет разрешить проблему нарушения принципов коммуникации: нет нужды (да и невозможно) соблюсти все принципы, важно только не нарушить те, которые для данного говорящего в данной коммуникативной ситуации оказываются первостепенными.

Хотя эта модель применима к порождению высказываний вообще, описание зачёркивания с её помощью играет особую роль потому, что это единственный случай, когда мы можем наблюдать соревнование кандидатов на практике: в обычной речевой ситуации оно практически ненаблюдаемо и может постулироваться только на основании интроспекции. Непреднамеренные речевые сбои с самоисправлением могли бы служить хорошим материалом для аналогичного исследования (тем более, что зачёркивание в интернет-коммуникации, по сути, является игровым изображением непреднамеренных самоисправлений), но подобный анализ малопродуктивен: чаще всего они не связаны с представлением нескольких точек зрения. Во-первых, в абсолютном большинстве самоисправлений в устной речи наблюдается семантический изоморфизм между забракованным фрагментом и его коррелятом-исправлением (в работе [Подлесская 2014], основанной на анализе корпуса из 40 устных рассказов, доля неизоморфных исправлений составляет 16,5%), тогда как в большинстве зачёркиваний общее между зачёркнутым и незачёркнутым фрагментами обычно не лежит на поверхности (пожалуй, максимальная близость — это использование литературного и жаргонного квазисинонимов или же гипероним и гипонима: *добропорядочные ~~люди~~ филологи*). Во-вторых, среди самоисправлений с изоморфизмом, по данным той же работы, почти половина всех коррекций является лексическими повторами, а среди коррекций-модификаций многие

являются исправлениями оговорок, фактологических ошибок, ошибок на лексическую сочетаемость и другими не связанными с семантикой и прагматикой исправлениями. Среди оставшихся коррекций можно найти семантико-прагматические самоисправления, для объяснения которых подходит предложенная в настоящей работе модель:

(7) *··· эээ ”” всякь= || ·· различные /–зме-еи,,*[8]

(8) *·· Назвали его И^{w}= ” || ·· \/–Ва-аня…*

Но, как нам кажется, непреднамеренность таких коррекций во многих случаях затрудняет объяснение выбора одного из двух кандидатов; кроме того, процент подобных самоисправлений по отношению к общему числу коррекций, как было показано выше, крайне мал. Из всего сказанного следует, что именно анализ преднамеренного письменного зачёркивания — этого, казалось бы, совсем частного и незначительного приёма языковой игры — позволяет пролить свет на общезначимые прагматические механизмы.

Однако возникает проблема: в каких случаях говорящий прибегает к выводу на поверхность двух кандидатов, а в каких ограничивается одним? Поскольку зачёркивание является приёмом языковой игры, ответ на этот вопрос лежит в плоскости более общего вопроса: зачем говорящий вообще использует любую языковую игру? Мы оставляем поиск ответа на этот глобальный вопрос исследователям общей теории юмора, но отметим, что при использовании приёма зачёркивания, как и в некоторых других видах языковой игры (в первую очередь, в каламбуре) одновременно высказывается несколько точек зрения на одну и ту же ситуацию, каждая из которых преследует какую-либо частную коммуникативную цель. Однако в случае каламбура говорящий ограничен выбором многозначных или омонимичных слов, а приём зачёркивания не накладывает на говорящего строгих формальных ограничений, тем самым, с одной стороны, позволяя пользоваться языковой игрой и людям, не столь мастерски владеющим словом, а с другой стороны — давая возможность применить языковую игру почти в любой ситуации. В свою очередь, применение языковой игры приводит к улучшению лица говорящего, поскольку подчеркивает его творческие способности. Эта возможность относительно просто улучшить лицо и привела к широкому распространению зачёркивания в блогосфере, предоставившей для этого необходимые технические средства.

8. Словесные варианты приёма зачёркивания

Не следует, тем не менее, относиться к зачёркиванию как к приёму, характерному исключительно для блогосферы. С блогами связан взрывной рост популярности зачёркивания, но семантический аналог зачёркивания

[8] О правилах нотации см. в [Кибрик, Подлесская (ред.) 2009].

использовался ещё в 90-х годах в Фидонете и IRC-сетях, поддерживающих только plain text. Написанная сразу за словом последовательность ^W обозначала, что предшествующее ей слово якобы не написано (соответственно, несколько ^W — при «зачёркивании» нескольких слов); аналогично использовалась последовательность ^H, обозначавшая «зачёркивание» одного символа. Данный графический приём восходит к соответствующим командам («удалить слово», «удалить символ»), использовавшимся в текстовых терминалах.

В современной интернет-коммуникации ^W в описанной функции очень редко, но всё же встречается:

(9) *Опоздала на самолет из-за аварии на ж/д, осталась в Англии навеки^W до среды* (из блога, 2013)

Как указывалось в [Пиперски, Сомин 2013], «в последнее время литуративы (зачёркивания — А. П., А. С.) становятся менее употребительными из-за того, что основные средства сегодняшней Интернет-коммуникации — социальные сети — не поддерживают даже простейшего форматирования текста». Однако пользователи, перешедшие в социальные сети из блогов, испытывают потребность в использовании привычного им приёма. Альтернативой графическому зачёркиванию является использование тегов <s>...</s> или <strike>...</strike>, которые не заменяются на зачёркнутый текст, как это произошло бы в блоге, но остаются написанными, а также, существенно чаще, само слово «зачёркнуто», заключённое в скобки (прямые, угловые или квадратные) или звёздочки; иногда оно используется в виде открывающего и закрывающего тега:

(10) *Ватикан не знает, как выкрутиться с гомосексуальными браками* (зачёркнуто) *с половыми извращениями* (комментарии на сайте, 2011)

(11) *хм...оказывается я удачно не* [зачёркнуто]*лоханулся*[/зачёркнуто] *прошился на «сумбур» от китайцев...* ☺ (твиттер, 2011)

Можно предположить, что этот приём продолжает распространяться уже и на тех пользователей, кто никогда не пользовался настоящим зачёркиванием.

В некоторых случаях слово *зачёркнуто* может модифицироваться наречиями:

(12) *Он утверждает, что его мои глаза очаровали. Наверняка обманывает, грудью *жирно зачёркнуто* сердцем чую :)* (из блога, 2013)

(13) *Ножки* (несильно зачеркнуто) *барды хороши!)))* (из блога, 2012)

(14) *Ты пишешь доносы *зачеркнуто* кляузы *яростно зачеркнуто* предупредительные записки администрации, если видишь «плохой» пост* (из блога, 2009)

Конечно, градации по силе зачёркивания плохо описываются теорией оптимальности, для которой победа одного кандидата над другим не имеет градаций,

однако в нашей модели теория оптимальности нужна лишь для предсказания, какой из фрагментов будет зачёркнут, а какой — нет, а уточнение способа зачёркивания, добавляющее в «одномерную» графическую игру новое измерение, по-видимому, не поддаётся формализации и обобщению. К тому же, например, в примере (14) не следует считать, что заняты три ступени пьедестала: незачёркнутое, зачёркнутое и яростно зачёркнутое. Кажется, что *доносы* и *кляузы* — в равной мере проигравшие кандидаты, а наречие *яростно* относится не к зачёркиванию конкретного слова, а ко всей ситуации, изображающей выбор подходящего слова и нарастающее раздражение пишущего при последовательных неудачах.

Встречаются и ситуации, где автор показывает своё намерение зачеркнуть, тем самым как бы осуществляя акт зачёркивания:

(15) *Сам поход в театр, с мужчиной, да на премьеру, да если еще и нарядиться подобающе, во всех своих бриллиантах и мехах (как тут зачеркивание делается? :))…* (из блога, 2014)

Наконец, появляется также специальный приём — отрицание зачёркивания, как в противовес предыдущему зачёркнутому тексту (16), так и отдельно (17–18):

(16) *Если вдруг решите, что ваша жизнь ~~бессмысленна и беспощадна~~ как то идиотична, посмотрите этот фильм, там по бесмысленности и беспощадности* (не зачеркнуто)*, а так же идиотичности мышления и поступков герои переплюнули всех.* (из блога, 2011)

(17) *Они дождались, пока мне исполнится 46 лет, пока я стану многодетной матерью и перейду в разряд к тому же матерей «одиноких», а также превращусь в старую развалину* (не зачеркнуто из принципа) (из блога, 2012).

(18) *все, я обиделась. *не зачеркнуто** (из блога, 2010)

В англоязычных текстах последних лет фиксируется тенденция к использованию слова *slash* с семантикой зачёркивания, где зачёркнутой считается часть фразы до слова *slash* (границы зачёркивания читатель должен определить сам; подробнее см. [Curzan 2013]):

(19) *I need to go home and write my essay slash take a nap.*
‘Мне надо пойти домой и написать эссе слеш вздремнуть’

В русскоязычных текстах подобное употребление не зафиксировано.

Отметим, что приём зачёркивания также использовался и в литературе задолго до появления интернета и блогов [Суховей 2008, гл. 3]. Так, графические зачёркивания очень часты в поэзии Нины Искренко, дебютировавшей в печати в конце 80-х годов [Орлицкий 2014]. Зачёркнутые слова в её стихах не учитываются размером:

(20) *Король королю короля с королем и оравой не глядя сдавал и сдавая мочился ~~постился~~ не глядя*

В поэтических текстах встречаются и вербальные зачёркивания, ср. (21)–(23):

(21) *коснуться — «бюст» зачеркиваю — уст!* (И. Бродский, «Двадцать сонетов к Марии Стюарт»)

(22) *Я глуп* (зачеркнуто)… *Я так неловок* (зачеркнуто)… *Я оскудел умом*. (Б. Ахмадулина. «Отрывок из маленькой поэмы о Пушкине»)

(23) *Весна!* (Зачеркнуто) *Прекрасный март…* (Зачеркнуто) *Голубоглазый март…* (Д. Самойлов. «Черновик»)

Впрочем, преднамеренным самоисправлением, похожим на интернет-зачёркивание, является только (21), а (22)–(23) скорее имитируют процесс порождения письменного текста.

9. Заключение

Таким образом, преднамеренное самоисправление в письменной речи — это весьма сложный приём и с точки зрения прагматики, и с точки зрения формальных средств выражения. Исследование этого приёма позволяет пролить свет как на общие прагматические механизмы коммуникации, так и на особенности письменной коммуникации в различных средах, в первую очередь в Интернете.

Литература

1. *Вахтин Н. Б., Головко Е. В.* 2004. Социолингвистика и социология языка. СПб.: ИЦ «Гуманитарная Академия»; Изд-во ЕУ в СПб.
2. *Грайс Г. П.* 1985. Логика и речевое общение. В кн.: Новое в зарубежной лингвистике. Вып. 16. Лингвистическая прагматика. М.: Прогресс. С. 217–237.
3. *Гусейнов Г. Ч.* 2008. Неполная коммуникация в блогосфере: эрративы и литуративы. http://www.speakrus.ru/gg/liturative.htm (проверено 15.02.2014)
4. *Занегина Н. Н.* 2009. Я этого не говорил: о литуративах, зачеркиваниях или мнимых текстах. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15). М.: Изд-во РГГУ. С. 112–115.
5. *Кибрик А. А., Подлесская В. И.* 2009. Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М.: Языки славянских культур.
6. *Кронгауз М. А.* 2013. Самоучитель Олбанского. М.: Астрель, Corpus.
7. *Орлицкий Ю. Б.* 2014. Что, почему и зачем зачёркивала Нина Искренко? // Доклад на семинаре «Зачёркивание как акт коммуникации» (СПб., 5–6 февраля 2014).

8. *Пиперски А. Ч., Сомин А. А.* 2013. Литуративы в русском интернете: семантика, синтаксис и технические особенности бытования. 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. Т. 1. С. 605–618
9. *Подлесская В. И.* 2014. Структура самоисправлений говорящего по данным корпуса устных рассказов // Язык. Константы. Переменные. Памяти Александра Евгеньевича Кибрика. СПб.: Алетейя. С. 96–104.
10. *Савицкая Л. С.* 2009. О приёме зачёркивания как средстве метакатегориальной организации модусной перспективы высказывания (на материале интернет-дневников). Вестник Нижегородского университета им. Н. И. Лобачевского. №6 (2). С. 346–349.
11. *Суховей Д. А.* 2008. Графика современной русской поэзии: диссертация … кандидата филологических наук. СПб.
12. *Brown, Penelope & Stephen Levinson.* 1978. Universals in language usage: Politeness phenomena. In Esther N. Goody, Questions and politeness. Cambridge: Cambridge University Press. Pp. 56–289.
13. *Brown, Penelope & Stephen Levinson.* 1987. Politeness: Some universals in language usage. Cambridge: Cambridge University Press.
14. *Curzan, Anne.* 2013. Slash: Not just a punctuation mark anymore. http://chronicle.com/blogs/linguafranca/2013/04/24/slash-not-just-a-punctuation-mark-anymore (проверено 15.02.2014).
15. *Dupriez, Bernard.* 1991. A dictionary of literary devices: Gradus, A–Z. Transl. & adapt. by Albert W. Halsall. Toronto & Buffalo: University of Toronto Press.
16. *Goffman, Erving.* 1955. On face-work: An analysis of ritual elements of social interaction. Psychiatry: Journal for the Study of Interpersonal Processes 18(3). 213–231.
17. *Grice, H. P.* 1975. Logic and conversation. In Peter Cole & Jerry L. Morgan (eds.), Syntax and semantics, vol. 3 New York. Pp. 41–58.
18. *Horn, Laurence R.* 2004. Implicature. In Laurence R. Horn & Gregory Ward (eds.), The Handbook of Pragmatics. Malden, MA & Oxford: Blackwell. Pp. 3–28.
19. *Kager, René.* 1999. Optimality Theory. Cambridge: Cambridge University Press.
20. *Prince, Alan & Paul Smolensky.* 1993. Optimality Theory: Constraint interaction in Generative Grammar. Ms.

References

1. *Brown P., Levinson S.* (1978), Universals in language usage: Politeness phenomena, in Questions and politeness, ed. by E. Goody, Cambridge University Press, Cambridge, pp. 56–289.
2. *Brown P., Levinson S.* (1987), Politeness: Some universals in language usage, CUP, Cambridge.
3. *Curzan A.* (2013), Slash: Not just a punctuation mark anymore, available at http://chronicle.com/blogs/linguafranca/2013/04/24/slash-not-just-a-punctuation-mark-anymore.

4. *Dupriez B.* (1991), A dictionary of literary devices: Gradus, A–Z, transl. & adapt. by Albert W. Halsall, University of Toronto Press, Toronto, Buffalo.
5. *Goffman E.* (1955), On face-work: An analysis of ritual elements of social interaction, in Psychiatry: Journal for the Study of Interpersonal Processes, vol. 18(3), pp. 213–231.
6. *Grice H. P.* (1975), Logic and conversation, in Syntax and semantics, vol. 3, ed. by P. Cole and J. L. Morgan, New York, pp. 41–58.
7. *Grice H. P.* (1985), Logic and conversation [Logika i rechevoe obshchenie], in New Trends in Foreign Linguistics. Vol. 16. Pragmatics in Linguistics [Novoe v zarubezhnoj lingvistike. Vyp. 16. Lingvisticheskaja pragmatika], Progress, Moscow, pp. 217–237.
8. *Guseinov G. Ch.* (2008), Incomplete communication in the blogosphere: erratives and lituratives [Nepolnaja kommunikatsija v blogosfere: èrrativy i lituraty], available at http://www.speakrus.ru/gg/liturative.htm
9. *Horn L. R.* (2004), Implicature, in The Handbook of Pragmatics, ed. by L. R. Horn and G. Ward, Blackwell, Malden, MA & Oxford, p. 4–28.
10. *Kager R.* (1999), Optimality Theory, Cambridge University Press, Cambridge.
11. *Kibrik A. A., Podlesskaja V. I.* (2009), Night Dream Stories: A corpus study of spoken Russian discourse [Rasskazy o snovidenijah: Korpusnoe issledovanie ustnogo russkogo diskursa], Jazyki skavjanskih kul'tur, Moscow.
12. *Krongauz M. A.* (2013), Teach yourself Olbanian [Samouchitel' Olbanskogo], Astrel', Corpus, Moscow.
13. *Orlitskij Ju. B.* (2014), What and whay did Nina Iskrenko cross out? [Chto, pochemu i zachem zachërkivala Nina Iskrenko], paper presented at the "Crossing out as a communication act" seminar (Saint Petersburg, 05–06.02.2014)
14. *Piperski A. Ch., Somin A. A.* (2013), Strikethrough on the Russian web: semantics, syntax and technical issues [Lituraty v russkom internete: semantika, sintaksis i tehnicheskie osobennosti bytovanija], in Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialog" (Bekasovo, May, 29th — June, 2nd 2013) [Komp'juternaja lingvistika i intellektual'nye tehnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferentsii "Dialog" (Bekasovo, 29 maja — 2 ijunja 2013)]. Vol. 12 (19), RGGU, Moscow, pp. 605–618.
15. *Podlesskaja V. I.* (2014), Structure of speaker's self-repairs based on a corpus of oral narratives [Struktura samoispravlenij govorjashchego po dannym korpusa ustnyh rasskazov], in Jazyk. Konstanty. Peremennye. Pamjati Aleksandra Evgen'evicha Kibrika [Language. Constants. Variables. In memoriam: Alexander Evgen'evich Kibrik], Aleteja, Saint-Petersburg.
16. *Prince A., Smolensky P.* (1993), Optimality Theory: Constraint interaction in Generative Grammar, ms.
17. *Savitskaja L. S.* (2009), Strikethrough as a means of metacategorical organization of the modus of an utterance in blogs [O priëme zachërkivanija kak sredstve metakategorial'noj organizatsii modusnoj perspektivy vyskazyvanija], in Vestnik Nizhegorodskogo universiteta im. N. I. Lobachevskogo, vol. 6 (2), Nizhnij Novgorod, pp. 346–349.

18. *Suhovej D. A.* (2008) The graphics of modern Russian poetry [Grafika soveremennoj russkoj poèzii], PhD thesis, Saint Petersburg.
19. *Vakhtin N. B., Golovko E. V.* (2004), Sociolinguistics and the sociology of language [Sociolingvistika i sociologija jazyka]. IC "Gumanitarnaja Akademija", izd-vo EU v SPb., Saint-Petersburg.
20. *Zanegina N. N.* (2009), I didn't say that: on liturатives, strikethrough, or spurious texts [Ja ètogo ne govoril: o liturativah, zachërkivanijah ili mnimyh tekstah], in Komp'juternaia lingvistika i intellektual'nye tehnologii: Po materialam mezhdunarodnoj konferentsii "Dialog 2009" [Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialog 2009"]. RGGU, Moscow, pp. 112–115.

«ТО ЕСТЬ, НЕ УБИЛИ, А ЗАРЕЗАЛИ САБЛЕЙ»: САМОИСПРАВЛЕНИЯ ГОВОРЯЩЕГО В УСТНЫХ РАССКАЗАХ

Подлесская В. И. (podlesskaya@ocrus.ru)

Российский государственный гуманитарный университет, Москва, Россия

На материале электронного корпуса устных рассказов предпринят качественный и количественный анализ самоисправлений говорящего, затрагивающих лексику, морфологию, синтаксическую структуру, сегментную фонетику и просодию. Показано, что в наиболее частотных паттернах самоисправлений, фрагмент, подлежащий коррекции, и его откорректированный коррелят являются формально и функционально изоморфными. Именно этот изоморфизм позволяет преодолеть последствия речевого затруднения.

Ключевые слова: устный дискурс; самоисправление; русский язык; корпус

“THEY SHOT HIM DEAD, OH, NO, THEY KNIFED HIM DEAD WITH A SABER”: SELF-REPAIRS IN ORAL STORIES

Podlesskaya V. I. (podlesskaya@ocrus.ru)

Russian State University for the Humanities, Moscow, Russia

The paper introduces a discourse oriented classification of repair types in Russian by addressing, inter alia, the following questions: (i) whether or not self-repairing entails speech disfluency; (ii) whether or not the fragment under repair and its repaired correlate are structurally isomorphic; (iii) does the speaker revise a lexical, a morpho-syntactic, or a phonologic shape of the reparandum. Basing on the data from the Prosodically Annotated Corpus of Spoken Russian, established classes of repairs were analyzed qualitatively and quantitatively. Fluent isomorphic repairs appeared to be the most frequent in the corpus, although fluent non-isomorphic repairs, as well as disfluent isomorphic and disfluent non-isomorphic repairs are also attested.

Keywords: spoken discourse; self-repair; Russian; corpus

1. Постановка вопроса

Задача работы — выявить языковые средства, которые задействует говорящий в тех случаях, когда обнаруживается несоответствие порожденного речевого фрагмента исходному замыслу[1]. Материалом послужили самоисправления, или (авто)коррекции, говорящего в четырех корпусах 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):

- NDS «Рассказы о сновидениях» (129 монологов, респонденты от 7 до 17 лет);
- SLS «Рассказы сибиряков о жизни» (17 монологов, респонденты от 19 до 70 лет);
- FLS «Веселые истории из жизни» (40 монологов, респонденты от 18 до 60 лет);
- SPS «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).

Для качественного и общего сравнительного количественного анализа использовались 817 эпизодов самоисправления, полученных методом сплошной выборки из всех четырех корпусов коллекции. Для детального количественного анализа использовались 194 эпизода самоисправления, зарегистрированных в FLS и 79 эпизодов самоисправления, зарегистрированных в SPS.

Основные принципы разметки дискурсивной информации в экспериментальных корпусах этой коллекции были сформулированы в работе [Кибрик, Подлесская (ред.) 2009]. В этой же работе были сформулированы базовые подходы к систематизации речевых сбоев и к их аннотированию. Для последующего изложения существенны следующие сведения о типах дискурсивной информации и способах их аннотирования:

- В транскриптах произведено деление на элементарные дискурсивные единицы (ЭДЕ). ЭДЕ — минимальный фрагмент дискурса, который способен вступать в смысловые отношения с другими фрагментами в структуре дискурса. Просодически, ЭДЕ — это автономная единица, интонационно организованная вокруг одного рематического (фразового) акцента; с синтаксической точки зрения ЭДЕ в сильной степени коррелирует с клаузой.
- Транскрипты снабжены просодической разметкой — в частности, указаны локализация акцентов и тональный тип акцента (нотируется перед словом иконически с помощью косых черт), размечены абсолютные и заполненные паузы. Ударный слог в слове — носителе фразового акцента подчеркивается.
- Точка прерывания (речевого сбоя) маркируется знаком «==» на границе ЭДЕ, и знаком «||» внутри ЭДЕ

[1] Работа выполнена при поддержке РГНФ, грант 12-04-00258

Дальнейшее изложение строится следующим образом. В разделе 2 вводятся два параметра классификации самоисправлений, которые позволяют описать, каким образом самоисправление встраивается в текущий дискурс; в разделе 3 один из этих параметров обсуждается более подробно; в разделе 4 рассматривается количественное распределение отдельных классов коррекций; в заключительном разделе 5 извлекаются уроки из полученных результатов.

2. Онлайн vs. оффлайн исправления; изоморфные vs. неизоморфные исправления

2.1. Трехчастная схема самоисправления

Прототипическими самоисправлениями, начиная с классических работ Levelt 1993, Shriberg 1994 и др., считаются самоисправления, строящиеся по трехчастной схеме «репарандум (фрагмент, подлежащий исправлению) / точка прерывания / репаранс (откорректированный коррелят)». В точке прерывания возможны заполненные паузы, лексические маркеры хезитации и прочие сигналы сбоя, ср.:

(1) *SLS*
··· эээ ””всякь= || ·· различные /–зме-еи,,,

(2) *FLS*
·· Поскольку-у \дверь ээ к тому моменту-у || ” \замок к тому моменту не / отмёрз,

2.2. Два свойства канонического трехчастного самоисправления

Прототипические трехчастные самоисправления обладают следующими двумя важнейшими свойствами. Во-первых, они строятся с использованием онлайн стратегии, при которой проблема устраняется сразу, как только она обнаружена, ср. (Levelt 1993: 478), или в уточненной формулировке (Finlayson et al.2011: 79–80) — «как только говорящий завершил перепланирование и готов продолжить артикуляцию». Плавное развертывание дискурса прерывается, наличие прерывания показывает говорящему, что репарандум следует устранить из текста, чтобы получить правильный с точки зрения говорящего фрагмент. Онлайн стратегия сопряжена с речевым сбоем, т. е. с нарушением лексико-грамматической и/или просодической когерентности дискурса, в том числе, с обрывом текущего фрагмента. По нашим корпусным данным в 40 % коррекций в точке прерывания имеется обрыв слова.

Во-вторых, для прототипических трехчастных самоисправлений характерен формальный и семантический изоморфизм фрагмента, подлежащего

исправлению, и его откорректированного коррелята. Ср. почти полную синонимию, тождество грамматического оформления и синтаксической функции репарандума *всяк[ие]* и репаранса *различные* в (1) или репарандума *дверь к тому моменту* и реперанса *замок к тому моменту в* (2).

2.3. Оба свойства канонического самоисправления не являются обязательными

Во-первых, не является единственно возможной онлайн стратегия. Говорящий может предпочесть оффлайн стратегию: исправление ошибки откладывается до завершения текущего речевого отрезка, и лишь после того, как этот отрезок транслирован слушающему, следуют эксплицитные признания ошибки и извинения по типу «О, простите, я высказался неверно/неточно». Оффлайн стратегия не предполагает нарушений плавного развертывания речи. Так, например, самоисправление в (3), вынесенное в заголовок статьи, и самоисправление в (4) полностью встроены в когерентный дискурс:

(3) *NDS*
/*Потом он один раз* вышел на /балкон,
·· и /его /–пристрели-или…
·· То есть не /пристрелили,
а-а \за= =\резали.
\Саблей.

(4) *SPS*
·· /–Утро^{h},,,
·· /мужик \–просыпается^{w}.
·· Даже не /мужик наверное,
/–парень,,,

Поскольку самоисправление при оффлайн стратегии встроено в плавно развертываемый дискурс, точка прерывания выпадает из трехчастной схемы. Ввиду отсутствия самоперебива и таких внешних сигналов сбоя, как например, обрывы слов, оффлайн самоисправления плохо поддаются автоматической аннотации; поэтому они редко исследуются на корпусных массивах (среди редких исключений — Maruyama, Sano 2006; Plug 2006).

Во-вторых, не является обязательным изоморфизм между фрагментом, подлежащим исправлению и его откорректированным коррелятом. Так, онлайн коррекции могут быть неизоморфными, если говорящий в принципе отказывается от забракованного фрагмента, изменяет исходный замысел и переходит к новому, когерентно построенному эпизоду. Фактически, в неизоморфных онлайн коррекциях репарандум есть, а репаранса — нет:

(5) *NDS*
····· И-и /\мы-ы с Саньком ·· тогда /пошли,
··· в \магазин,
чего-нибудь \покупать,
·· а-а ==
···(0.5) и Са= ==
и я \проснулˆся,
на /\самом там-м ·· интересном –месте.

(6) *FLS*
У нас был ·· **чёрный** потом огромный /потолок ==
\А!,
ещё \больше того,
он же стал это \тушить.

2.4. Независимость двух параметров

Два параметра самоисправлений — онлайн vs. оффлайн и изоморфные vs. неизоморфные — являются независимыми, и в реальном дискурсе наблюдаются все четыре возможных комбинации их значений.

В каноническом трехчастном самоисправлении, примеры (1)–(2), представлена комбинация «+ онлайн, + изоморфизм». В примерах (5)–(6) представлены неизоморфные онлайн коррекции, т. е. комбинация «+ онлайн, — изоморфизм». Дальше всего от канонической трехчастной трехчастной схемы отстоят неизоморфные оффлайн коррекции («– онлайн, — изоморфизм») — в них самоисправление встроено в полностью когерентный дискурс, нет точки прерывания, откорректированный коррелят (а иногда и забракованный фрагмент) представлен диффузно в протяженном фрагменте дискурса. Так, в примере (7) первая ЭДЕ произносится с падающим тоном во фразовом акценте, что сигнализирует завершенность; исправление добавлено пост-фактум и сигнализируется дискурсивным маркером *в смысле (что)*:

(7) *FLS*
·· Но-о /он кстати был очень \приличный молодой человек.
Ну в смысле что он был /молодой,
и-и ·· /\ну-у ·· как бы /внешне не \выглядел как фрик.
Единственное что был \голый.

Самоисправления этого рода стоят в одном ряду с такими дискурсивно когерентными построениями, как постпозитивные уточнения, дополнения и проч. Внутри этого класса самоисправления могут выделяться лишь по семантическому критерию, т. е. лишь в том случае, когда говорящий объявляет «недействительным» какой-то фрагмент транслированного значения. В примере (7) говорящий понял, что употребил слово *приличный* неточно: он хотел

сказать, что молодой человек, если не считать отсутствия одежды, ВЫГЛЯДЕЛ прилично, но не имел в виду квалифицировать его поведение как поведение приличного (т. е. соблюдающего нормы) человека. Часто грань между исправлением и уточнением провести довольно трудно. Так, в (8) и (9) фрагмент, следующий после дискурсивного маркера *то есть*, следует, скорее, квалифицировать как уточнение/дополнение, поскольку говорящий не отменяет ничего из сказанного:

(8) *FLS*
Ну мы пошли купили билет на тот же самый /самол<u>ё</u>т,
··· стоило это раза в три /дор<u>о</u>же,
то есть за эту сумму можно было туда-обратно и ещё раз /туд<u>а</u> слетать,

(9) *FLS*
И он должен был поехать в \Ташк<u>е</u>нт.
··· {ЧМОКАНЬЕ} И там ··· торчать какое-то \вр<u>е</u>мя.
·· {ЧМОКАНЬЕ} \В<u>о</u>т.
То есть у него был /докл<u>а</u>д,
а потом он ещё должен был ··· какое-то время там \остав<u>а</u>ться.

Наконец, последняя из четырех возможных комбинаций — оффлайн изоморфные коррекции («– онлайн, + изоморфизм») — представлена в примерах (10)–(11), где откорректированный коррелят появляется после когерентного и полностью завершенного фрагмента с падающим тоном во фразовом акценте. При этом морфологическое оформление коррелята полностью изоморфно забракованному фрагменту. Например, в (10) форма винительного падежа единственного падежа слова *коробку* идентична форме забракованного фрагмента *посылку* и лицензируется глаголом *возьму*. Точно также в (11) откорректированный коррелят *кошки* дублирует подлежащее *коты*, хотя и вынесено за пределы интонационно и грамматически цельнооформленной ЭДЕ:

(10) *FLS*
<НРЗБ> «В-вам как \пос<u>ы</u>лку нужно оформить.»
Я говорю
«Какую /–пос<u>ы</u>лку?,
где я вам возьму \пос<u>ы</u>лку?!
Вот это || ·· –к-кор<u>о</u>бку?»,

(11) *FLS*
ну у нас там у \многих жили кстати в общаге-е ээ \кот<u>ы</u>.
·· \К<u>о</u>шки.

Изоморфизм в оффлайн коррекциях может проявляться и опосредованно — например, через встраивание откорректированного коррелята в конструкцию, изоморфную конструкции с забракованным фрагментом. Так,

в (3)–(4) откорректированный коррелят вводится через изоморфную конструкцию с эксплицитным отрицанием; а в примере (12) происходит замена обстоятельственного слота в структурно идентичных ЭДЕ:

(12) *NDS*
··· Я вечером /\уснул,
·· просыпаюсь \вечером.
···· Днём /\уснул,
а /–вечером ··· \просыпаюсь.

Изоморфизм репарандума и репаранса — одно из наиболее эффективных средств, позволяющих говорящему поддержать структурную целостность текста, нарушенную сбоем, а слушающему — помочь восстановить исходный замысел говорящего. Этот параметр будет более подробно исследован в следующем разделе.

3. Типы изоморфизма между забракованным фрагментом и откорректированным коррелятом

3.1. Повторы и модификации

Внутри изоморфных самоисправлений выделяется два основных типа — повторы и модификации.

Максимально полно изоморфизм репарандума и репаранса проявляется при повторах. Повторы могут быть вызваны следующими функциональными причинами:

- говорящий пытается выиграть время для планирования (хезитация);
- говорящий не удовлетворен тем, что начал произносить, пытается подобрать более удачный вариант, не справляется, решает за неимением лучшего довершить начатое;
- говорящий понимает, что приступил к фрагменту слишком рано, добавляет недостающую информацию и возобновляет фрагмент

В (13) приведен пример онлайн повтора, а в (14) — пример частичного офлайн повтора:

(13) *FLS*
···· Начал ==
··· Просёк эту /фишку,
·· и начал ==
они ж коты \чего делают,
они ·· –писают,
···· начал описывать им ··· –тапочки например какие-нибудь,,,

(14) *FLS*
·· И один раз когда мы с сестрой /спа̲ли,
·· (Кстати конечно я тоже этого не /по̲мню,
нам рассказала /\ба̲бушка с мамой.
\Да̲.)
···· когда мы /спа̲ли,
·· наша **бабушка** нав= || ·· намыла **столько** /–винограда,

Частичный изоморфизм репарандума и репаранса представлен в модификациях — репарандум и репаранс относятся к одному семантическому и\или грамматическому классу и претендуют на один и тот же слот в грамматической или, шире, дискурсивной структуре, но противопоставлены по некоторому параметру, ради которого и осуществляется коррекция. Наиболее распространенный тип модификации — лексическая. Лексическая онлайн модификация демонстрируется в (15); лексическая оффлайн модификация, маркируемая контрастной просодией, представлена выше в (12), а также в (16) и (17):

(15) *NDS*
·· Назвали его И^{w}= ” || ·· \/–Ва̲-а̲ня…

(16) *FLS*
·· И вот папа ··· шестого числа ·· значит /доложи̲л,
··· {ЦОКАНЬЕ} ·· \пятого числа папа /доложи̲л,
··· и шестого пошёл бегать ·· ээ по всяким /авиака̲ссам,

(17) *SPS*
··–Сел на /–лы̲жи,,,
·· {ЦОКАНЬЕ} ·· \вста̲л на лыжи,
···· /и \↑по̲ехал!

Лексические модификации в общем случае удовлетворяют требованию Правильной Структуры (“Well-formedness Rule”, Levelt 1993: 485–489): забракованный фрагмент и его откорректированный коррелят входят в отношение близкое к альтернативному сочинению, и в частности, они правильным образом могут быть встроены в соответствующий альтернативный вопрос, ср. «Назвали Иван или Ваня?»; «Доложил пятого или шестого?»; «Сел на лыжи или встал на лыжи?». Однако даже для лексических модификаций подобного рода вопросы не всегда адекватны. Трудности возникают, например, при коррекции референциальных выражений, см. (18)–(19) («*кто-то или какой-то мужчина бежит?»; «*это какой-то или некий знак?»:

(18) *NDS*
···· И я /смотрю̲,
кто-то’ ==
какой-то \мужчи̲на за нами бежит.

(19) *FLS*
··· и тут Брентон ···пробегает мимо \/меня-а,
···· с таким ·· **конусом**.
·· Ну вот которые \ставят там-м ” для чего-то на /дорогу,
то есть это какой-то= || некий \знак,

Еще менее применимым оказывается требование Правильной структуры к модификациям, затрагивающим грамматику и фонологию, которым посвящен следующий раздел.

3.2. Изоморфные самоисправления, затрагивающие грамматику и фонологию

3.2.1. Морфологические модификации

Изоморфные модификации внутрисловно выражаемых категорий обычно требуют повторного воспроизведения слова:

(20) *SLS*
····на правом берегу реки Воронеж ···· находился || находилась маленькая станция \Отрожка.

(21) *SPS*
У его /жены ··· /скоро должно было случиться день \рождения.
·· –\Был случиться.

В редких случаях однако обнаруживаются мофрфологические модификации без повторного воспроизведения слова, что необычно для флективного фузионного языка (ср.обсуждение в Wouk 2005), ср. замену основы без необходимой замены алломорфа приставки:

(22) *NDS*
···· /Потом^{h} ·· ’ ···· подо’= ·· =/бежала ко мне \розоваяw ·· мышка^{w},

3.2.2. Модификации с использованием согласуемых препаративных подстановок

Особый класс модификаций в русском языке составляют самоисправления, с использованием прономинальных выражений, проецирующих грамматическую форму отложенной составляющей (английский термин placeholders, см. Podlesskaya 2010), ср. мужской род репарандума и репаранса в (23) и женский в (24):

(23) *NDS*
А /вы ещё нас /поведёте в \этот ||·· в тренажёрный \зал?

(24) *FLS*
··· и-и ···· \Шороха мы держали ·· на-а этой ||·· на /пан<u>е</u>ли,,,

3.2.3. Модификации сегментного фонологического состава

В этот класс отнесены оговорки per se — модификации, при которых забракованный вариант представляет собой звуковую цепочку, не являющуюся фактическим словом языка:

(25) *FLS*
и они значит вмесело || весело маршируют \нав<u>е</u>рх.

3.2.4. Модификации просодии

В редких случаях говорящий может исправлять просодическую конфигурацию репаранса. Так, в (26) предпоследняя строка была первоначально замыслена как незаключительная и произнесена, соответственно, с подъемом тона, однако в результате автокоррекции строка повторена в качестве заключительной с падением тона во фразовом акценте:

(26) *NDS*
···· а с= || ··· /двойн<u>я</u>шки,
···· /они’ ·· /\’**ого<u>р</u>а<u>ж</u>ивали**,
···· там где \пос<u>а</u>жено,
ставили /к<u>о</u>лышки ==
···· ставили \к<u>о</u>лышки^{h},

3.2.5. Конструкционные модификации

При самоисправлениях этого класса репарандум подвергается структурным изменениям, которые делают невозможным встраивание откорректированного коррелята в исходную структуру (изменения порядка слов, замена именной группы на глагольную и др.). Вместе с тем, «остатки» исходной структуры сохраняются, что позволяет считать эти коррекции изоморфными. Так в (27) представлено эксплицитное отрицание забракованного фрагмента, а в (28) — замена первоначально задуманного именного сказуемого (*больной?/ хромой?/ несчастный?*) на глагольную клаузу:

(27) *FLS*
·· обслуживали быто= || ” ну всякую ==
не /бытов<u>у</u>ю,
а-а ·· \копиров<u>а</u>льную технику.

(28) *SPS*

и /он с перелом-м= =мом \руки-/ноги,

··· теперь такой-й ==

” короче на /лыжах он теперь больше не \будет кататься.

4. Количественное распределение самоисправлений

Таблица 1 демонстрирует частотность самоисправлений в четырех корпусах коллекции «Корпуса звучащей речи» (PrACS-Russ). Зарегистрированная частотность самоисправлений — 2,3–3,4 в минуту, 1,8–2,9 на 100 слов — хорошо согласуется с данными по другим языкам. В целом, частотность самоисправлений обычно фиксируется в примерном интервале от одного до семи в минуту; в диалогах — выше, чем в монологах, 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 and Watson 2008); для венгерских диалогов — 3,8 в минуту (Németh 2012); для японских монологов — 1,2 на 100 слов (Maruyama and Sano 2006). Таким образом, полученные нами данные позволяют предположить, что межъязыковое, межжанровое и межвозрастное варьирование частотности самоисправлений находится в пределах интервала, отражающего универсальные тенденции речепроизводства. При этом тексты корпуса SPS, отличающиеся от трех остальных корпусов по жанру — это рассказы по картинкам, а не «истории из жизни» — по-видимому, демонстрируют свойства меньшей спонтанности, так как в них частота самоисправлений ниже, чем в трех других корпусах:

Таблица 1. Общее число и частотность самоисправлений в исследованных корпусах

Корпус	Число слов	Время звучания (мин)	Общее число испр	Испр/мин	Испр/100 слов
SLS	5000	40	132	3,3	2,6
NDS	14000	120	412	3,4	2,9
FLS	7000	70	194	2,8	2,8
SPS	4500	35	79	2,3	1,8

По-видимому, именно меньшая спонтанность, приводит к тому, в корпусе SPS выше, чем в трех других корпусах, доля онлайн коррекций, хотя, в целом, по всем корпусам доля онлайн коррекций систематически значительно превосходит долю оффлайн коррекций, см. Таблицу 2:

Таблица 2. Распределение онлайн и оффлайн самоисправлений в исследованных корпусах

Корпус	Общее число испр	Онлайн (# / %)	Оффлайн (# / %)
SLS	132	108 / 82 %	24 / 18 %
NDS	412	343 / 83 %	69 / 17 %
FLS	194	164 / 85 %	30 / 15 %
SPS	79	71 / 90 %	8 / 10 %

Более частные типы самоисправлений исследовались на материале двух корпусов «Веселые истории из жизни», FLS, и «Истории о подарках и катании на лыжах», SPS. Таблицы 3а и 3б показывают, что «золотым стандартом» являются онлайн изоморфные самоисправления. В корпусе FLS 146 онлайн изоморфных коррекций составляют 75 % от общего числа (194) и 89 % от числа онлайн коррекций (164); в корпусе SPS 64 онлайн изоморфные коррекции составляют 81 % от общего числа коррекций (79) и 90 % от числа онлайн коррекций (71):

Таблица 3а. Распределение основных типов самоисправлений в корпусе FLS

	Онлайн	Оффлайн	Всего
Изоморфные	146	16	162
Неизоморфные	18	14	32
Всего	164	30	194

Таблица 3б. Распределение основных типов самоисправлений в корпусе SPS

	Онлайн (дисфлу)	Оффлайн (флу)	Всего
Изоморфные	64	8	72
Неизоморфные	7	–	7
Всего	71	8	79

Таблицы 4а и 4б показывают, что среди изоморфных самоисправлений число модификаций незначительно превышает число повторов — в FLS соотношение 45,9 % : 54,1 %, в SPS соотношение 40,6 % : 59,4 %. Среди модификаций наиболее частотными являются лексические и конструкционные; самыми редкими — просодические, хотя частные классы изоморфных модификаций демонстрируют больший разброс между двумя корпусами, чем более общие типы самоисправлений. Можно предположить, что и здесь оказывает влияние жанр текстов, но это нуждается в более тщательной проверке.

Показательно распределение частот оговорок в исследованных корпусах. В корпусе FLS 7000 слов, следовательно, 15 оговорок в корпусе дают долю 1,7 на 1000 слов. В корпусе SPS 4500 слов, следовательно, 5 оговорок в корпусе дают долю 1,1 на 1000 слов. Снижение доли оговорок в рассказах по картинкам — еще одно проявление меньшей спонтанности таких текстов по сравнению в рассказами из жизни. Частотность оговорок оказалась сопоставима с данными, которые имеются в литературе по другим языкам (1–2 на 1000 слов), ср. Garnhem et al. (1981) по данным Лондонско-Лундского корпуса — 1 оговорка на 1000 слов в английских диалогах; Eklund (2004: 258ff.) по данным корпуса шведских диалогов (включая диалоги человека с человеком и человека с компьютером) — от 1,5 до 2,4 на 1000 слов. Это позволяет предположить, что частотность оговорки отражает действие общих механизмов речепорождения, и в меньшей степени связана с лингвоспецифическими и жанроспецифическими факторами. Однако — это предположение, как и все другие приведенные соображения о количественном распределении речевых сбоев, безусловно, нуждаются в проверке на более объемных выборках, сбалансированных по языкам, возрасту говорящих, жанрам и другим параметрам.

Таблица 4а. Распределение изоморфных самоисправлений в корпусе FLS

Тип коррекции		**Онлайн**	**Оффлайн**	**Всего**
Повторы		67	8	75
Модификации	лексические	31	6	39
	морфологические	4	–	3
	с использованием согласуемых препаративных подстановок	11	–	11
	сегментного фонологического состава (оговорки)	15	–	15
	просодические	2	–	2
	конструкционные	16	2	17
	Итого:	79	8	87
Всего		146	16	162

Таблица 4б. Распределение изоморфных самоисправлений в корпусе SPS

Тип коррекции		**Онлайн**	**Оффлайн**	**Всего**
Повторы		26	1	27
Модификации	лексические	10	3	12
	морфологические	7	2	8
	с использованием согласуемых препаративных подстановок	–	–	–
	сегментного фонологического состава (оговорки)	5	–	7
	просодические	–	–	–2
	конструкционные	16	2	18
	Итого:	38	7	45
Всего		64	8	72

5. Выводы

- Были рассмотрены самоисправления говорящего в электронной коллекции «Рассказы о сновидениях и другие корпуса звучащей речи» — первом открытом ресурсе с систематически размеченной фразовой сегментацией, просодией, речевыми сбоями, паузацией и другими явлениями спонтанной устной речи.
- Была предложена классификация самоисправлений, позволяющая оценить способ встраивания самоисправления в текущий дискурс.
- Параметры предложенной классификации ориентированы на следующие исследовательские вопросы: (а) связано ли самоисправление с нарушением структурной целостности текущего дискурса; (б) являются ли забракованный фрагмент и его откорректированный коррелят структурно изоморфными; (в) затрагивает ли самоисправление лексику, морфологию, синтаксис, сегментную или супрасегментную фонологию забракованного фрагмента.
- Было показано, что говорящие отдают существенное предпочтение онлайн изоморфным самоисправлениям по сравнению со всеми остальными классами, зарегистрированными в исследованных корпусах. Можно предположить, что данный способ для говорящего оказывается наименее трудозатратным, а слушающему позволяет наиболее точно реконструировать исходный замысел говорящего. Однако это предположение нуждается в проверке с привлечением не только корпусных, но и экспериментальных методов.

Литература

1. *Кибрик А. А., Подлесская В. И.* (Ред.). Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М.: ЯСК, 2009.

References

1. *Kibrik, A. A. and V. I. Podlesskaya* (eds.): 2009, Rasskazy o snovidenijax: korpusnoe issledovanie usntogo russkogo diskursa [Night Dream Stories: A corpus study of spoken Russian discourse]. Moskva: Jazyki Slavjanskix Kul'tur.
2. *Eklund, Robert*: 2004, Disfluency in Swedish human–human and human–machine travel booking dialogues. PhD thesis, Linköping Studies in Science and Technology, Dissertation No. 882, Department of Computer and Information Science, Linköping University, Sweden.
3. *Finlayson* et al.: 2011, 'Fluency or accuracy: What matters when correcting errors in spoken dialogue?', in Ian Finlayson, Robin Lickley, and Martin Corley (eds.), Architectures and mechanisms of language Processing (AMLaP 2011), Centre Universitaire Paris Descartes, 79–80.
4. *Fraundorf, S. H. and D. G. Watson:* 2008, 'Dimensions of variation in disfluency production in discourse', in J. Ginzburg, P. Healey, and Y. Sato (eds.), Proceedings of LONDIAL 2008, the 12th Workshop on the Semantics and Pragmatics of Dialogue, London: King's College London, 131–138.
5. *Garnham, Alan, Richard C. Shillcock, Gordon D. A. Brown, Andrew I. D. Mill and Anne Cutler*: 1981. 'Slips of the tongue in the london-lund corpus of spontaneous conversation', Linguistics 19(7–8), 805–818.
6. *Levelt, Willem J. M.*: 1983, 'Monitoring and Self-Repair in Speech', Cognition 14. 41–104.
7. *Levelt, Willem J. M.*: 1993, Speaking: From Intention to Articulation [ACL-MIT Series in Natural Language Processing].
8. *Maruyama, Takehiko and Sano, Shin'ichiro*: 2006, 'Classification and Annotation of Self-Repairs in Japanese Spontaneous Monologues', in LPSS — Linguistic Patterns in Spontaneous Speech, Taipei, November 2006, 283–298.
9. *Németh, Zsuzsanna*: 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.
10. *Plug, L.*: 2006, 'Speed and reduction in postpositioned self-initiated self-repair', York Papers in Linguistics 2(6). 143–162.
11. *Podlesskaya, Vera*: 2010, 'Parameters for typological variation of placeholders', in N. Amiridze, Boid H. Davis and Margaret Maclagan (eds), Fillers, Pauses and Placeholders. [Typological Studies in language (TSL), vol. 93]. Amsterdam/Philadelphia: John Benjamins, 11–32.
12. *Shriberg, Elizabeth Ellen*: 1994, Preliminaries to a Theory of Speech Disfluencies. PhD thesis, University of California, Berkeley.
13. *Spokencorpora*: 2013, «Рассказы о сновидениях и другие корпуса звучащей речи». Prosodically Annotated Corpus of Spoken Russian. Pilot version. Online: http://spokencorpora.ru
14. *Tseng, S.-C.*: 2006, 'Repairs in Mandarin conversation', Journal of Chinese Linguistics 34(1), 80–120.
15. *Wouk, Fay*: 2005, 'The syntax of repair in Indonesian', Discourse Studies 7 (2), 237–258.

ОПЫТ АНАФОРИЧЕСКОЙ РАЗМЕТКИ КОРПУСА И РАЗРЕШЕНИЯ АНАФОРЫ С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ

Протопопова Е. В. (protoev@gmail.com),
Бодрова А. А. (anastasie.bodrova@gmail.com),
Вольская С. А. (svetlana.volskaya@gmail.com),
Крылова И. В. (krylova93@gmail.com),
Чучунков А. С. (scarywound@gmail.com)
Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Алексеева С. В. (alexeeva@opencorpora.org),
Бочаров В. В. (bocharov@opencorpora.org),
Грановский Д. В. (granovsky@opencorpora.org)
Проект «Открытый Корпус», OpenCorpora.org

Ключевые слова: корпуса текстов, краудсорсинг, разрешение анафоры, синтаксическая разметка

ANAPHORIC ANNOTATION AND CORPUS-BASED ANAPHORA RESOLUTION: AN EXPERIMENT

Protopopova E. V. (protoev@gmail.com),
Bodrova A. A. (anastasie.bodrova@gmail.com),
Volskaya S. A. (svetlana.volskaya@gmail.com),
Krylova I. V. (krylova93@gmail.com),
Chuchunkov A. S. (scarywound@gmail.com)
Saint Petersburg State University, St. Petersburg,Russia

Alexeeva S. V. (alexeeva@opencorpora.org),
Bocharov V. V. (bocharov@opencorpora.org),
Granovsky D. V. (granovsky@opencorpora.org)
OpenCorpora.org

The paper describes the noun phase and anaphora annotation in OpenCorpora and compares it to that in other corpora. We discuss the choice of representative texts for anaphoric annotation and the basic principles of syntactic annotation. In case of noun phrase annotation we followed the scheme introduced earlier for morphological annotation: it was carried out in two stages: firstly, all noun phrases and some other syntactic units were annotated by a heterogenous group of people, then a linguist compared all markup results and found the best one, or corrected mistakes. We present some annotation results and cases of annotator's disagreement and proceed to introduce our data-driven anaphora resolution system based on decision trees. We then list the features used to fit the classificator and discuss their relevance and some changes which improved the classificator performance. We also present out rule-based approach to automated noun phrase extraction using Tomita parser. A baseline for anaphora resolution is introduced and we compare it with our results.

Keywords: anaphora resolution, corpora, crowdsourcing, syntactic annotation

1. Introduction

The task of anaphora and coreference resolution is quite important in automated text processing and understanding, since these are often used to maintain text coherence. Facing this task state-of-the-art coreference resolution systems exploit various supervised machine learning techniques [13] and thus require manually annotated data. Our goal is therefore to build such a corpus for Russian as a part of OpenCorpora project[1]. Moreover, no such corpus for Russian is freely available now and that is perhaps why the number of papers on using machine learning techniques for Russian coreference resolution is rather low.

We start with annotating noun phrases in a part of our corpus including some news articles and short stories. Groups of words acting as one syntactic unit, which contain nouns, (such as complex preposition *в связи с 'in view of'*) were also annotated. We tried to make our annotation scheme as simple as possible because some annotators are not linguists. A small portion of texts was annotated by several people, the others being given only to one person. A linguist then compared (in case of several annotations) or corrected the resulting annotation. Then a text was annotated with anaphoric relations for 3rd person pronouns, some demonstrative pronouns in anaphoric use, possessive pronouns and relative pronouns (*который 'which / that'*, *кто 'who'* etc.). An annotator can only mark the relation between a selected pronoun and previously annotated NP, this is why the previous stage is important. We later discuss the problem concerned with choosing the right phrase in case there are several referring to one entity.

One of the ways to evaluate such a corpus is to use it for the task of anaphora resolution. We build a classifier which uses some morphological and textual features. The corpus annotated for anaphora resolution track was used as training data. A tool for automated noun phrase extraction was implemented using Tomita-parser[2]. With its help positive and negative training examples were then generated.

---

[1] OpenCorpora, available at: http://opencorpora.org/

[2] Tomita-parser, available at: http://api.yandex.ru/tomita/

2. Related Work

Two tasks may be distinguished in our work: first of all, we describe our experience in anaphora annotation, then, we present the anaphora resolution system. It seems natural to mentions previous work made in these two fields in two separate subsections.

2.1. Anaphoric Annotation in Corpus

A great number of annotation schemes were proposed for anaphoric and coreferential annotation. The most famous is MUC scheme[3] which aims at high inter-annotator agreement and was used in MUC competitions. The annotation rules are quite simple; personal and possessive pronouns, names and other named entities, bare nouns as a part of coreference chain are markables (should be annotated). As for relations, they annotate basic coreference (two NP refer to the same object), bound anaphors, apposition and some other more specific cases. Another significant resource for English was developed as a part of Ontonotes corpus [2]. An important difference from MUC (ACE) scheme concerns the annotation of verbal phrases: in Ontonotes such phrases may be marked as antecedents or coreferring phrases (There is an example of predicative anaphora in section 3.3). Anaphoric and bridging (or associative) relations were also annotated in the ARRAU corpus [9], [10]. In general, they followed MUC scheme, but propose another markup format based on TEI instructions[4].

As for Slavonic languages, an annotation scheme different from previous ones was proposed for Prague Dependency Treebank (PDT) [7]. All referential entities including generic and abstract ones were subject to annotation. Predicate nominals and appositions were not considered as corefent. Textual coreference including pronominal one is annotated along with bridging relations.

2.2. Anaphora Resolution Using Machine Learning Techniques

Many anaphora resolution systems proposed in last 50 years are rule-based. They are, however, worth concerning because their rules use so called anaphora resolution factors. Perhaps an exhaustive survey of such systems is presented in [6]. A recent highly appreciated system is Stanford Coreference Resolution system [12]. We will now focus on some important works regarding these factors as well as machine learning approach to anaphora resolution.

[13] propose a corpus-based system, which learns from small amount of data using quite a restricted number of features. The vector for each pair of markables consists of the following 12 factors: distance in sentences, whether each markable is a pronoun, whether markables are equal, whether NP is definite or demonstrative,

---

[3] MUC-7 Coreference Task Definition, available at: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/co_task.html

[4] TEI:Text Encoding Initiative, available at: http://www.tei-c.org/index.xml

agreement in number, gender and semantic class (each counted separately), alias (i.e. if NP is another name for the same entity) and appositive features. C5 learning algorithm (based on decision tree) was used to learn from this data. The evaluation was conducted on MUC sets, the best result reported achieved F-measure of 62.6% and precision 58.6%. They also analyzed classification errors. The improvements to this system were proposed in [8], which lead to F-measure of 70.4%. They used 53 features but then reduced this number to 41, including various syntactic and semantic features.

Different machine learning techniques were tested in [5].Using previously examined factors of referential choice [3], [4], they achieved quite a high accuracy of anaphora resolution up to 88.7%.

3. Corpus Annotation

3.1. Texts for Annotation

Our chief aim was to create a representative corpus, which may be then used as a training data for anaphora resolution systems, so texts for annotation were chosen with special attention. First of all, we considered genre structure of this subcorpus. We consider press materials to be the most characteristic of modern language and in particular of such phenomena as anaphora and coreference, that is why a half of our subcorpus is composed of news articles. Another half is made up of fiction texts, blog posts and encyclopedic texts. Then the texts were filtered automatically: we examine news size and choose texts of its average size for all listed genres. Then they were reviewed manually and were filtered again. We exclude texts which do not include many examples of anaphoric relations. About a third of texts were filtered out and were substituted by other texts with the highest number of pronouns.

The total size of corpus planned was about 100,000 words. By the time, however, only 18,000 are annotated.

3.2. Noun Phrase Annotation

As mentioned above, we started with annotating noun phrases and some more specific units. The following kinds of phrases are subject to annotation: basic noun phrases, names and named entities, pronouns, complex conjunctions and prepositions, parenthetical expressions and prepositional phrases. The exhaustive list of groups is presented in table 1.

Each annotator chooses a text and annotate it sentence by sentence. The annotation process may be divided into the following steps: the annotator first finds all nouns in the phrase and then marks all simple groups (1–7). Basic noun phrases may include adjectives, ordinal numbers, adverbs (очень 'very') and particles (не 'not'). An annotator should mark group as proper name if it contains a proper noun. Thus, in the following expression 'Марина Павловна Трубецкая' 'Marina Pavlovna Trubetskaja' three groups of the second

kind should be annotated. Groups 4–7 are specified by the lists from Russian National Corpus (RNC)[5]. Complex groups (8–15) should include at least two simple groups.

Table 1

	Group	Example
1	basic noun phrase	*не очень интересный журнал*
2	proper name	*прекрасную Францию*
3	numeral	*сто двадцать пять*
4	complex preposition	*в течение*
5	adverbial expression	*без оглядки*
6	parenthetical expression	*к слову сказать*
7	complex conjunction	*до тех пор пока*
8	complex proper names	*Марина Павловна Трубецкая*
9	proper name with generic term	*княжна Трубецкая*
10	apposition	*статья 112*
11	prepositional phrase	*от меня*
12	coordinated NPs	*Маша и Петя*
13	complex noun phrase (NP containing two or more NPs)	*куртка Маши*
14	numeral phrase	*три яблока*
15	complex pronoun	*друг друга*

An annotator should also mark heads for those phrases where it is not obvious. We consider it to be obvious in cases where head can be easily found automatically: basic NPs, prepositional phrases, enumeration. We also introduce special tags ALL and NONE for enumerations and groups 4–7 respectively.

When all sentences of the text are annotated, the mark-up should be revised. A moderator reviews the annotation sentence by sentence and can accept annotated groups or mark their own.

3.3. Anaphoric Annotation

For the anaphoric annotation the pronouns from the list were highlighted in text and all annotators can mark relations between these pronouns and preceding NPs. The annotation follows several rules: first of all, we agreed to mark the relation between a pronoun and its nearest member of coreferential chain. Thus, in the following sentence, the relation between 'Фернандо Алонсо' 'Fernando Alonso' and 'свой' 'his' should be annotated (1):

(1) ***Фернандо Алонсо*** *в первый раз в* ***своей*** *карьере пилота Формулы-1 выиграл Гран-при Монако.* ***'Fernando Alonso*** *won Grand-Prix Monaco for the first time in* ***his*** *Formula 1 driver career'*

[5] http://ruscorpora.ru

Moreover, the antecedent should be the maximal possible group. We do not annotate predicative anaphora such as (2):

(2) ***Шёл дождь. Это*** *нас остановило.* ***'It was raining. This*** *stopped us.'*

We do not annotate cataphora though we have seen several examples of it in texts such as:

Хотя ***он*** *казался спящим,* ***Иван*** *думал.*
'Although ***he*** *seemed to be sleeping,* ***John*** *was thinking.'*

One reason for this is that we would like to limit classifier's search space and the number of possible antecedents in text.

4. Anaphora Resolution System

We implemented a data-driven anaphora resolution system, which relies on previously annotated corpus. The pairs of markables in corpus are deduced automatically by a special tool for noun phrase extraction and the training vectors are computed for all possible pairs 'antecedent—anaphora'. Pairs are marked as positive/negative examples and then are used to fit the classifier. These stages are described in corresponding subsections.

4.1. Noun Phrase Extraction

To extract all possible markables equivalent to those used in manual annotation we developed a NP extraction tool using Tomita-parser[6]. Originally a tool for fact extraction, Tomita deals with context-free grammars and key-word dictionaries. For the current purpose, a grammar for NP extraction was used to process sentences. For each rule, the parser finds the longest substring meeting the requirements. Thus, our groups were defined in terms of sequences of tags. Sometimes, our restrictions were insufficient and the rules were corrected many times. An XML output of parser was then combined with the information from our tokenizer and a markable was represented as a pair of identifiers—text id and token id. Precision and recall are 0.81 and 0.82 respectively.

4.2. Feature Vectors

A set of features is necessary for a classifier to define whether a pair is bound with anaphoric relation or not. Our features are based on practical as well as theoretical conclusions and are meanwhile easy to compute. All extra information was obtained through open-source tools and resources.

---

[6] Tomita-parser, available at: http://api.yandex.ru/tomita/

We divide our features into three groups: lineal, morphological and syntactic features. These classes are described below. Each feature was computed for anaphor and its possible antecedent head.

Lineal Features

1. The number of proper nouns between anaphor and antecedent
2. The number of sentences between anaphor and antecedent
3. The number of potential anaphors for the given antecedent between given anaphor and given antecedent
4. The number of nouns between anaphor and antecedent
5. The number of anaphoric pronouns between anaphor and antecedent
6. The number of possible antecedents for the given anaphora between given anaphor and given antecedent

Morphological features

These features were computed using our morphological mark-up (OpenCorpora morphological dictionary) and no disambiguation was carried out.

1. Part-of-speech of the antecedent
2. Whether antecedent is in nominative
3. The number of verbs in the sentence containing antecedent
4. The number of nouns in the sentence containing antecedent
5. The number of conjunctions and pronominal adjectives in the sentence containing antecedent
6. The number of nonfinite verb forms in the sentence containing antecedent

Syntactic features

The syntactic information for these features was obtained with the help of MaltParser[7].

1. Whether antecedent is subject
2. Whether anaphor is subject

4.3. Classifier

Our learning method is based on decision trees and follows the ID3 algorithm [11]. Test pairs were extracted from documents as it was preciously done for training pairs. The vectors were post-processed, because the result on the data as is was very low (18% accuracy). The following steps were undertaken:

1. Binarize all lineal features to features 'is more than', 'is between X and Y' etc.
2. Remove all pairs where no positive pair is found for a pronoun.
3. Treat all examples where antecedent is too far from anaphor as negative.
4. Add feature counts from the nearest possible antecedent to all possible antecedents for given pronoun.

[7] MaltParser, available at: http://www.maltparser.org/

The classifier starts from the anaphor and proceeds till a positive pair is found. Then it works till the first negative example and marks all further pairs as non-anaphoric.

5. Evaluation

In this section we would like to present the results of manual annotation as well as the results of automated anaphora resolution.

5.1. Manual Annotation

Although at first there were controversial opinions on the annotation principles, the annotation itself seems to be quite simple for annotators. Seven annotators participated in this task and 9,100 groups (5,788 simple and 3,312 complex) were annotated. First of all, we can observe annotator-moderator agreement. The annotators make mistakes only in difficult cases (such as the order of combining units into complex groups) though they are no professional linguists.

We use two metrics to estimate inter-annotators' agreement: Cohen's kappa [15] and F-mean. They show quite good results for pairs of annotators: for simple groups kappa varies from 0.61 to 0.97 and F-mean is more than 0.9. The results concerning complex groups are somewhat lower: best kappa scores vary from 0.67 to 0.75. These figures suggests that the annotation manual was clear for all annotators and that the task itself is not very difficult.

5.2. Anaphora Resolution

Our baseline system marks as anaphoric pair pronoun and the nearest possible antecedent. The accuracy is computed in the following way: a pair is marked correctly as anaphoric if its antecedent's head equals to that of reference antecedent. Baseline accuracy is therefore about 50.4% on the corpus of 94 documents and somewhat more than 2,000 pairs.

The current system was built using the corrections mentioned above and achieved the accuracy of 52.04%. Here we present a part of the tree (1 is for antecedent and 2 is for anaphor) (3):

(3)
```
'number of NPRO',
{
 '>4' => [
    'POS 2',
    {
     '2_3' => 'no',
     '3' => [
              'number of nonfinites for 2',
```

```
{
 '>10' => 'no',
 '<undef>' => 'no',
 '1' => 'no',
 '2' => [
          '1 is nominative',
          {
           '1' => [
           'number of nouns <= 3 for 1',
           {
            '1' => 'no',
            '0' => 'yes'
           }
```

6. Conclusion

In this paper we have described our attempt to create a corpus with anaphoric annotation and an anaphora resolution system. Here we would like to outline some of the future directions. First of all, we have seen that the part of NP annotation is time-consuming so it may be conducted in semi-automated way as we have already implemented a tool for fully automated NP extraction. The anaphora resolution system may be improved with many additional features and, furthermore, be transformed into coreference resolution system. On the other hand, we can pay more attention to the training data with respect to proportion of positive and negative examples and more complicated learning algorithms.

References

1. *Bocharov V. V., Alexeeva S. V., Granovsky D. V., Protopopova E. V., Stepanova M. E., Surikov A. V.* (2013) Crowdsourcing morphological annotation [Morfologicheskaja razmetka korpusa silami volontërov], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2013” [Komp’iuternaia Lingvistika i Intellektual’nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii “Dialog 2013”], Bekasovo, pp. 109–115
2. *Hovy E., Marcus M., Palmer M., Ramshaw L., Weischedel R.* (2006), OntoNotes: The 90% Solution, available at: http://bbn.com/resources/pdf/HLT-NAACL-2006-OntoNotes.pdf
3. *Kibrik A. A., Dobrov G. B., Zalmanov D. A., Linnik A. S.* (2010), Referential choice as a multi-factor probabilistic process [Referentsial’nyj vybor kak mnogofaktornyj verojatnostnyj protsess], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2010” [Komp’iuternaia Lingvistika i Intellektual’nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii “Dialog 2010”], Bekasovo, pp. 173–181

4. *Kibrik A. A.* (2011), Reference in Discourse, Oxford Studies in Typology and Linguistic Theory
5. *Kibrik A. A., Linnik A. S., Dobrov G. B., Khudyakova M. V.* (2012), Optimizing a machine learning base model of referential choice [Optimizatsija modeli referentsial'nogo vybora, osnovannoj na mashinnom obuchenii], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2012"], Bekasovo, pp. 237–247
6. *Mitkov R.* (1999), Anaphora resolution: the state of the art, available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.29.6235&rep=rep1&type=pdf
7. *Nedoluzhko A., Mírovský J., Ocelák R., Pergler J.* (2009), Extended Coreferential Relations and Bridging Anaphora in the Prague Dependency Treebank, available at: http://ufal.mff.cuni.cz/~nedoluzko/koref_anot/DAARC_Nedoluzhko.pdf
8. *Ng V., Cardie C.* (2002), Improving Machine Learning Approaches to Coreference Resolution, available at: http://acl.ldc.upenn.edu/acl2002/MAIN/pdfs/Main347.pdf
9. *Poesio M.* (2004), The MATE/GNOME Proposals for Anaphoric Annotation, Revisited, available at: http://acl.ldc.upenn.edu/W/W04/W04-2327.pdf
10. *Poesio M., Artstein R.* (2008), Anaphoric Annotation in the ARRAU Corpus, available at: http://catalog.ldc.upenn.edu/docs/LDC2013T22/lrec08_297.pdf
11. *Quinlan J. R.* (1986), Induction of Decision Trees. Machine Learning, Vol. 1, I. 1. available at: http://www.dmi.unict.it/~apulvirenti/agd/Qui86.pdf
12. *Recasens M., De Marneffe M., Potts C.* (2013), The Life and Death of Discourse Entities: Identifying Singleton Mentions. In Proceedings of NAACL-HLT 2013.
13. *Soon W. M., Ng H. T., Lim D. C. Y.* (2001), A Machine Learning Approach to Coreference Resolution of Noun Phrases, available at http://acl.ldc.upenn.edu/J/J01/J01-4004.pdf
14. *Lee H., Chang A., Peirsman Y., Chambers N., Surdeanu M., Jurafsky D.* (2013), Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules. Computational Linguistics, Vol. 39, N. 4. MIT Press, Cambridge, MA.
15. *Cohen, Jacob* (1960), A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20 (1).

RECENT ADVANCES IN (DEEP) REPRESENTATION LEARNING

Schütze H. (hs0711@cis.uni-muenchen.de)

University of Munich, Germany

Traditionally, natural language processing (NLP) systems have made use of resources compiled by (computational) linguists based on linguistic theory that provide rich information about linguistic objects. For example, computational lexica specify morphological paradigms and subcategorization frames of verbs. In contrast, statistical NLP systems frequently start out with no explicit representation of linguistic objects and instead learn what they need from training data on a task-by-task basis. A third approach—which has gained much interest recently—is to learn generic representations of linguistic objects and then reuse them for a wide variety of tasks. Its premise is that giving an NLP system non-task-specific generic information about words and other linguistic objects will help it in performing well at a particular task.

Examples of such generic representation models include the vector space model, dimensionality reduction, clustering and deep learning. I will review recent research results in representation learning and discuss benefits and drawbacks of the three approaches.

О КЛАССЕ РУССКИХ ПАРАМЕТРИЧЕСКИХ НАРЕЧИЙ

Семенова С. Ю. (sonya_sem@mail.ru)

ИНИОН РАН, Российский государственный гуманитарный университет, Москва, Россия

Ключевые слова: русская параметрическая лексика, параметрическая информация, количественное наречие, частеречное соответствие, ближайшая этимология, коннотация количества, полисемия

ON THE CLASS OF RUSSIAN PARAMETRIC ADVERBS

Semenova S. Ju. (sonya_sem@mail.ru)

INION RAS, Russian State University for the Humanities, Moscow, Russia

The paper deals with Russian parametric adverbs i.e. those revealing the values of the quantitative parameters (*gluboko* [deeply], *chasto* [often / thickly / frequently], *redko* [rarely / seldom], *bystro* [rapidly / quickly], *izdaleka* [from afar] etc). Characteristics of parametric adverbs seem to be much less investigated (in particular, in the perspective of information extraction) than those of parametric nouns, adjectives, and verbs. A number of grammatical and semantic groups of adverbs are presented. The parametric meaning is found to be distributed among various traditional grammatical and semantic classes of adverbs. For parametric adverbs morphologically derived from adjectives, we discuss semantic priority or lack thereof with respect to adjectives. The parametric meaning can take place for a secondary sense of an adverb, so that ambiguity, connotations, and implication are essential in the descriptions aimed at information extraction. The correspondence between the quantitative meaning of the adverb and the name of the physical value (*izdaleka — rasstojanie* [distance]) are considered. Corpora examples of various types of parametric data coded with the help of parametric adverbs are presented.

Key words: Russian parametric words, parametric data, quantitative adverb, part of grammar correspondences, nearest etymology, connotation of quantity, ambiguity of sense

Среди полнозначных частеречных классов русской количественной параметрической лексики, представляющей интерес как для теоретических исследований, так и для прикладных задач (понимание текста, извлечение данных и знаний), наиболее изученными являются параметрические существительные (или имена количественных параметров: *высота*, *глубина*, *емкость*, *скорость*, *температура*, *количество* и др.) и прилагательные (*высокий*, *низкий*, *глубокий*, *вместительный*, *легкий* и др.). Менее изучены (как отдельный, прагматически значимый подкласс) параметрические глаголы, образующие весьма размытый и в семантическом, и в формально-грамматическом отношении пласт глагольной лексики (от классических, «ядерных» параметрических глаголов *весить*, *стоить*, *вмещать* и нек. др., относительно регулярным образом мотивирующих параметрическое имя (по типу морфологической деривации пара *весить — вес* несколько отличается от других: имя есть результат усечения), до глаголов различных групп, со «встроенным» большим или малым значением параметра, выступающим как компонент семантической структуры: *обморозить, кипеть, недозреть*, *похудеть*, *мчаться* и др.). Параметрическая лексика названных частеречных категорий, прежде всего пространственная, рассматривалась в русистике в целом ряде аспектов: для существительных, в частности, описывалась полисемия [2,3, 15], изучались психолингвистические особенности номинации величин [3]; для прилагательных тоже исследовались психолингвистические свойства [8, 11]; для глаголов — поведение под оператором отрицания [2], соответствие с номинативной параметрической лексикой [5, 13]. Доказал свою плодотворность и типологический подход, затрагивающий, в частности, адъективную параметрическую лексику [9].

При этом, насколько нам известно, в семантических исследованиях практически не рассматривались как отдельная категория параметрические наречия (т. е. слова наподобие *высоко, поздно, скоро, близко, издалека*, *подолгу* и т. п.), соответственно, они не получили и детального систематизированного описания, ориентированного на извлечение параметрической информации. Между тем, такие наречия в качестве представителей параметрической лексики интересны и с теоретической, и с практической точки зрения. Так, в теоретическом плане любопытно сопоставление пар «прилагательное — наречие», выражающих значение того или иного параметра, с точки зрения ближайшей этимологии (см. ниже подклассы 1.1–1.3); также требуют рассмотрения соотношение между наречием и названием параметра (*издалека — расстояние*); количественные коннотации исходно непараметрических наречий (например, кванторных: *поминутно, кое-когда*); полисемия параметрических наречий (*глубоко перекапывать почву — глубоко чувствовать*); синтаксические конструкции с наречиями, релевантные для поиска параметрической информации. Перечисленные аспекты (кроме первого, этимологического) существенны и в практическом отношении, поскольку связаны с задачей извлечения.

Конечно, русским наречиям как большому грамматическому классу, насчитывающему целый ряд групп со своими грамматико-семантическими свойствами, посвящена значительная литература; среди источников следует, в частности, указать [1, 7, 10, 17–19]; библиография может быть расширена. Но сквозь «призму» information extraction наречия детально не рассматривались.

В данной работе, носящей предварительный, эскизный характер, хотелось бы наметить пути к преодолению этой лакуны, по крайней мере, обозначить и охарактеризовать с указанной точки зрения основные группы параметрической адвербиальной лексики.

Прежде всего, о терминологии. К параметрическим будем относить наречия, выражающие значения количественных параметров. Вообще, денотатами терминологического элемента «параметрический» в отечественной лексической семантике, в силу сложившейся традиции, стали языковые единицы двух типов, в зависимости от логической роли слова в высказывании, связанном с параметрической информацией: параметрическое слово обозначает либо саму величину, либо ее значение. В общем смысле, параметрическое слово — это слово, соотнесенное с «параметром» (или признаком, обладающим некоторым [переменным] значением); чаще всего рассматриваются параметры физические, но могут быть также экономические и абстрактные математические (*длина*, *стоимость*, *тангенс*). Параметрические существительные (параметрические имена) обозначают величины; кроме того, к параметрическим традиционно относят также имена, обозначающие качественные признаки объектов — *цвет*, *диагноз* и др. [4] (В принципе, к параметрическим можно было бы относить и существительные, выражающие количественные значения на компонентном уровне: *великан*, *юноша, пятиэтажка*, *стометровка* и т. п., но это не соответствует терминологической традиции). К параметрическим глаголам относят в первую очередь глаголы, обозначающие величины; ядерные параметрические глаголы связаны с именами величин трансформационным образом (*весить* — *вес*, *насчитываться* — *численность* и т. п.); параметрическими также считаются глаголы с компонентом «значение величины» (*мчаться* — *большая скорость*). Параметрические прилагательные — это прилагательные разных структурных моделей, выражающие значения величин: *высокий*, *холодный*, *пятиметровый*, *нулевой*, *сверхзвуковой* и др. И параметрические наречия будем определять по аналогии с прилагательными; они, тоже образованные в соответствии с разными структурными моделями, выражают значения величин.

Укажем ряд групп параметрических наречий. Как будет видно далее, количественные параметрические значения (или по крайней мере, коннотации) встречаются у наречий (или их лексем) практически всех традиционно выделяемых структурных и семантических групп (среди наречий на *-о (-ски)*, отыменных и местоименных, наречий места, времени, образа действия). Признак «параметрический» прочерчивает дополнительный семантический наречный класс, как бы поверх традиционного деления в грамматиках.

1. Вначале о параметрических наречиях, относящихся к самому продуктивному морфологическому типу (на *-о* и *-ски*). Среди слов, образованных по этой модели, немало традиционных наречий образа действия. (Наличие у ряда наречий образа действия параметрических значений отмечено в [18: 114]).

Взгляд на наречия с «параметрической» позиции способствует различению среди них трех групп, связанных с ближайшей этимологией наречия — двух более-менее четких групп и одной «остаточной».

1.1. Наречия на *-о (-ски)*, образованные от тех параметрических (по большей части качественных) прилагательных, которые исходно определяют сущности предметных (а не ситуативных /процессных, событийных/ категорий): *высоко, низко*, *широко, узко*, *глубоко*, *толсто, гигантски*, *тяжело, холодно*, *густо*, *людно* и др. (ср. *высокое дерево*, *широкий стол*, *тяжелый камень*, *глубокая река, толстый слой*, *гигантский кит, людная улица* [определяемые объекты — предметы]; *холодная вода*, *густой мед* [объекты — вещества]). Соответственно, наречия в исходных значениях (т. е. не в производных, не в метонимических или метафорических) выражают большую / малую величину параметров, у которых предметна (в широком смысле, включая вещества, вместилища, географические объекты, существ и др.) исходная таксономическая категория измеряемых объектов [13]: для параметров *высота/ ширина/глубина (Х –а)* характеризуемый объект Х — изначально ПРЕДМЕТ (для *глубины* — такая разновидность предметов, как ВМЕСТИЛИЩЕ), для *температуры* и *вязкости* — Х в первую очередь ВЕЩЕСТВО. Для подобных наречий, морфологически образованных от широко понимаемых «предметных» прилагательных, не подлежит сомнению не только морфологическая, но и семантическая производность, это в полном смысле «отадъективные» наречия.

Подчеркнем, что в данной работе в рассмотрение взяты только исходные значения наречий, например, пространственные: *провода протянуты высоко*; *волосы коротко пострижены*. Метафорические значения наречий, которые обусловливают изменение природы параметров [15] (например, наречия *длинно* и *коротко* как оценки по темпоральным параметрам: *длинно / коротко излагать*), требуют отдельного рассмотрения.

Оговоримся также, что мы рассматриваем только ближайшую этимологию, на формальном уровне связанную лишь с частеречными различиями прилагательного и наречия. Если рассматривать этимологию более дальнюю, когнитивная картина усложняется: например, пара *тяжелый-тяжело* восходит к глаголу *тянуть* [«Этимологический словарь русского языка» М. Фасмера], а глагольная лексика больше ассоциируется с наречиями, чем с прилагательными: *тяжелый* — такой, который *тянет*, а значит, его *тяжело* нести.

Отметим и то, что название параметра не всегда морфологически связано с парой «прилагательное — наречие» и не всегда очевидно для наречия; так, слова *холодно* и *температура* морфологически не родственны, а для слова *густо* наиболее подходящим параметром может выступать *концентрация*, *плотность, интенсивность, количество*: *густо заваренный сироп* (параметр — *концентрация*); «*Места вдаль дороги густо заселены…*» ([НКРЯ]; параметр — *плотность* (метафоризированная); *Не густо!* (параметр — *количество*) и т. п.

Для задач извлечения название параметра/признака и параметрическая лексема, способная выражать значение, должны быть связаны в словарном описании. В [12] описывается примерная структура краткой статьи наречия в прикладном формализованном семантическом словаре, нацеленном, в том числе, на задачи извлечения. Для параметрических наречий название релевантного параметра (или, по крайней мере, тематической группы параметров) там предложено размещать в зоне энциклопедической информации.

1.2. Наречия на *-о (-ски)*, морфологически соотнесенные с параметрическими прилагательными, но представляющиеся первичными с точки зрения семантики, что, очевидно, обусловлено исходной ситуативной (процессной, событийной) категорией характеризуемого прилагательным (и соответствующим параметром [14]) объекта: *близко*, *быстро, выгодно, громко*, *давно*, *далеко, дистанционно, долго, катастрофически, метко*, *надежно, недавно, однократно, подробно*, *поздно, размашисто*, *рано*, *ритмично, стремительно, эпизодически* и др. Соответствующие прилагательные можно толковать, опираясь на наречия, т. е. примерно по такой схеме: *быстрый* — «такой, который движется или происходит быстро»; *близкий* — «такой, который расположен близко»; *громкий* — «такой, который звучит громко», *меткий* (выстрел) — «такой, который произведен метко (характеристика по параметру *точность*)» и т. п.

Надо сказать, что в грамматических описаниях не указывается семантическая первичность таких наречий; например, в [17] все наречия на *–о (-ски)* описываются как отадъективные. Аргументом в пользу безусловной производности наречий в парах «прилагательное — наречие» может служить наличие усложняющих морфологическую структуру наречного слова словообразовательных элементов в других языках: англ. *–ly*, франц. *–ment*, итал. *–mente*; ср. *loud — loudly, bruyant — bruyamment, periodico — periodicamente*. (При этом в английском языке возможна и наречная форма с нулевым суффиксом, для которой примат прилагательного неочевиден: *Don't talk so loud*.)

В [18] отмечается семантически нетривиальный характер соотношения «прилагательное — наречие», в частности обращено внимание на несоответствие грамматических типов прилагательных и наречий, выделенных в [7]: «…обстоятельственному наречию *далеко* соответствует не относительное, а качественное прилагательное *далекий*» [18: 102]. Правда, с одной стороны, указанное несоответствие носит, скорее, условный, терминологический характер, с другой стороны, это несоответствие в цитируемой работе отмечено и для пары (*высокий — высоко*), в которой наречие производно.

В толковых словарях для наречий на *–о (-ски)* обычно не создаются самостоятельные статьи; статьи наречий либо отсутствуют, либо являются отсылочными (ср., например, фрагмент статьи прилагательного *эпизодический*, в которой наречие упомянуто как дериват: «*эпизодический* — 1. предпринимаемый от случая к случаю, не систематический…2. несущественный, не имеющий большого значения…3. появляющийся только в отдельных эпизодах…; *Эпизодически,* нареч.» [Большой толковый словарь РЯ — СПб., 1998]). Однако если исходить из примата наречий, выражающих значение «ситуативных» параметров, целесообразно описывать такие наречия в отдельных статьях. (Конечно, в прикладных задачах в этом нет необходимости; например, в [12] отмечается путь генерации статей наречий из статей прилагательных в автоматизированном режиме. Но в рамках традиционной семантической лексикографии отдельное описание представляется оправданным с теоретической точки зрения).

При этом семантический приоритет наречия перед прилагательным можно усмотреть не только среди параметрических количественных наречий,

но и среди значительного круга других наречий образа действия: *безапелляционно — безапелляционный, методично — методичный* и т. п.

1.3. Наречия на *-о* (*-ски*), для которых не просматривается четкого семантического приоритета ни у одного из членов пары «прилагательное — наречие»: *жарко, капитально, круто, часто* и др. В самом деле, прилагательное *частый* обычно характеризует события (*частые респираторные заболевания* и т. п.), и потому может рассматриваться как вторичное по отношению к наречию. С другой стороны, исходное значение прилагательного является пространственным, с предметной таксономией участника (*частый лес, чаща*), и потому тут имеется конкуренция интерпретаций. Для таких наречий, на наш взгляд, тоже целесообразны отдельные словарные описания.

Укажем параметрические наречия некоторых других структурных моделей.

2. В кодировании параметрической информации участвуют традиционные наречия меры и степени: *очень, немного, слишком* и др.; основные дискурсивные наречия малой степени изучены в [1]; сравнительные свойства членов синонимического ряда «слишком» описаны в [10]. Наречия меры и степени являются количественными в широком смысле; но собственно параметрическую информацию они выражают лишь в определенных контекстах; например, если наречие *слишком* определяет параметрическое прилагательное, наречие или слова *много* и *мало*: *Тесто слишком соленое*; *Ты слишком редко проверяешь оборудование* (примеры из [10]); «*… если иностранных банков станет слишком много, то хорошего в этом будет мало*» (А. Крашаков [НКРЯ]), или параметрический глагол: *Он слишком похудел*. Употребления в контексте слов, обозначающих неизмеряемые свойства / ситуации (*слишком заносчивый*; *слишком устал*) мы не считаем параметрическими. Как интенсификаторы параметрических прилагательных и наречий, наречия меры и степени выражают, так сказать, производную параметрическую информацию, «оценку оценки»; как интенсификаторы параметрических глаголов с семантикой изменения количества они (например, при эллиптическом опущении наречия *сильно*) способны характеризовать собственно величину изменения: *Слишком выросла безработица* (параметр — *рост2*). Для наречий малой степени круг собственно параметрических употреблений также является более узким, чем общих спектр допустимых контекстов: *еды еле хватило*; *едва слышный* и др. (примеры из [1]; параметры — *количество*; *громкость*).

Как известно, в этом подклассе есть пересечения с другими подклассами, например, меру и степень обозначает ряд наречий на -о (-ски): *значительно, исключительно* в том числе с метафорическим значениями [16]: *фантастически, головокружительно* и т. п.: «*За десять лет, что мы в контакте, Эндрю фантастически разбогател*» (С. Довлатов[НКРЯ]).

3. Релевантны наречия, образованные от числительных и семантически близких к ним слов: *вдвое, вдвоем, единожды, дважды, троекратно, попарно, вничью* (наречие местоименного происхождения, выражающее значение

параметра *счет*), *ежегодно* (значение частоты/кратности) и т. п.: *«Ежегодно по итогам учебного года школа проводит летние лагеря-семинары»* (объявление [НКРЯ]).

Параметрические значения с коннотацией большой частоты имеют наречия, соответствующие квантору общности и образованные от мелких единиц измерения времени *поминутно*, *ежесекундно* и т. п.: *Радиотелескопы ежесекундно принимают огромное количество радиосигналов со всей Вселенной….»* (А. Латкин [НКРЯ]).

4. Отыменные и отадъективные наречия с параметрическими значениями, на формальном уровне представляющие собой застывшие предложно-падежные формы: *изредка*, *издали, издалека*, *подолгу*, *допоздна*, *спозаранок*, *поблизости*, *рядом*, *годами*, *ввысь* и др.: «*Большой опыт и у бортинженера Бударина, он дважды подолгу работал на станции «Мир»* (С. Лесков [НКРЯ]). Для таких «сращений», в той или иной мере сохраняющих смысл предлогов и грамматических форм и потому являющихся семантически более сложными, чем «простые» исходные лексические пары типа *далеко-далекий*, задача собственно параметрической интерпретации наречия, с целью извлечения информации, зачастую обедняет заложенный в слово смысл; так для наречия *допоздна* релевантен прежде всего компонент «большая продолжительность», а остальные нюансы ситуации, характеризуемой наречием — то, что действие, скорее всего, происходит вечером, а ночью или утром прекращается (хотя возможно и более редкое употребление *спать допоздна*, т. е. до позднего времени утром), может игнорироваться в прикладных задачах.

5. Некоторые наречия отрицательной семантики, в том числе наречия с приставкой *без-*, имеющие значение (или коннотацию) большого или малого количества: *беззвучно* — очень тихо, *бесконечно* — очень много (или мало, ср. *бесконечно малая величина*), или долго, или сильно, *беспрерывно* — очень часто, *безграмотно* — с большим количеством ошибок (или с грубыми ошибками), *безлюдно* — нет или мало людей (или они не видны [6]), *бессрочно* — на большой срок, *безотказно* (о работе механизма) — очень надежно, *бесснежно* — нет или мало снега.

В словах, указывающих на «актуальную» бесконечность (*бесконечно*, *бессрочно*, *безотказно*), проступает идеализированная модель действительности, согласно которой определенные сущности, ограниченные по времени и пространстве, мыслятся как бесконечные и вечные.

Собственно, буквальное значение отсутствия («нулевое» значение) также выражает параметрическую информацию: *бесплатно*, *безденежно, безвозмездно, беспошлинно* (буквальные обозначения нулевого размера выплат), *беспосадочно* (нулевое количество посадок) и др.

Нулевое количество, которое предстает в языке как выделенное, маркированное значение параметров, может обозначаться и бесприставочными наречиями: *порожняком*, *пусто, даром, вхолостую, молча*.

6. Параметрическую информацию способны выражать и местоименные наречия, у которых отмечаются количественные коннотации: *как-то* (в темпоральном значении) — однократно или небольшое число раз; ср. «*Вы как-то сказали, что у вас нет любимых фильмов…*» [НКРЯ]; *когда-то* (в значении «в прошлом») — давно: «*Когда-то он служил в королевской армии…*» (В. Быков [НКРЯ]); при этом кратность самого действия в далеком прошлом могла быть и большой: *он когда-то обожал ходить за грибами*; *когда-нибудь* (в значении «в будущем») — нескоро: «*Если я когда-нибудь соберусь обзавестись постоянным домом…*» (Ю. Пешкова [НКРЯ]); *кое-когда* — редко: «*Так-то, всурьез, я не курю, а кое-когда балуюсь*» (М. Шолохов [НКРЯ]) и т. п. Малые значения обусловлены коннотацией малого количества у квантора существования.

7. Определенное место среди параметрических наречий занимают композиты разных моделей: *крупномасштабно*, *сверхурочно*, *круглосуточно*, *быстротечно*, *скоропостижно*, *стопроцентно* и др. Для них тоже закономерен вопрос о примате наречия над прилагательным; например, для пары *скоропостижно — скоропостижный* он очевиден. При этом, наречия соответствуют далеко не всем составным прилагательным; например, для адъективной модели «число+единица измерения» образование наречий практически невозможно: **трехметрово*; модель реализуется лишь для такой периферийной единицы, как *кратность* (нечто среднее между единицей измерения и параметрическим именем): *однократно*, *десятикратно* и т. п.; единица измерения *процент* не дает продуктивной модели: *стопроцентно* (параметр — *вероятность*), но *?пятипроцентно*. Заметим, что генерация реально не существующих наречий является одним из приемов языковой игры: *Торгово-развлекательный центр… пущай народ торгово развлекается!*

8. Параметрическую информацию могут выражать и наречия, являющиеся дериватами причастий: *умеренно, отдаленно* (в буквальном пространственном значении; *отдаленно напоминать* — метафоризированное непараметрическое употребление), *ускоренно*, *щадяще* и др.): «*Трубы в наростах ржавели,…отдаленно рокотала вода*» (В. Астафьев [НКРЯ]); «*Процесс старения происходит ускоренно…*» (Д. Гранин [НКРЯ]); « [диета]…*действует щадящее — вы худеете с оптимальной для организма скоростью…*» [НКРЯ].

Конечно, приведенный материал имеет предварительный характер; прежде всего, ставилась задача обозначить основные группы наречных лексем, способных выражать информацию о количественных величинах, показать некоторые связи с более изученными классами параметрической лексики (прилагательными, существительными).

Необходимо отметить принципиальную незамкнутость класса параметрических наречий. Количественные значения проявляются весьма разнообразно и «размыто» — в форме семантических компонентов разной степени значимости для разных слов, коннотаций, импликатур. Так, среди «бытовых» наречий образа действия (их представительный перечень имеется в [18]) немало имеющих отношение к параметрам тех или иных тематических групп.

Коннотации большой скорости (темпа) имеются у наречий *торопливо*, *ходко*, *проворно*, *расторопно*, *стремглав* и др., а у наречий *тягуче*, *чинно*, *потихоньку*, *пешком*, *шагом* и др. в той или иной мере воплощается смысл «медленно». Или, на периферию средств выражения длительности попадают наречия *сезонно* (значит, недолго), *подробно, обстоятельно* (следовательно, долго).

Требуют исчисления случаи «депараметризации» по типу [15], т. е. утраты наречием параметрического значения, главным образом за счет механизма метафоры (например, наречие *тяжеловесно* в «дальнем» переносном смысле утратило значение большого веса: …*"Зимин сожалел, что действовал столь тяжеловесно, можно было просто арендовать маленькое помещение"* (Г. Горелик [НКРЯ]). Наречия развивают и неметафорические значения, не имеющие при этом связи с исходным параметром; скажем, в сочетании *плотно прикрыть дверь* речь не идет о параметре *плотность*.

По сути дела, каждая указанная группа (как и каждое полисемичное наречное слово) требует отдельного, объемного изучения. Кроме полисемии, соответствия конкретным величинам, большой интерес представляет и синтаксис распространенных обстоятельствами предложений, несущих параметрическую информацию.

Автор благодарит студентов 4 курса Института лингвистики РГГУ — слушателей спецкурса по русской параметрической лексике [16], участвовавших в сборе лексики и в обсуждении интересных случаев; особенно хочется отметить Е. Иншакову, И. Смирнову, Л. Зисера.

Литература

1. *Апресян В. Ю.* Семантика и ее рефлексы у наречий усилия и малой степени // Вопросы языкознания. — 1997, № 5. — С. 16–34.
2. *Апресян Ю. Д.* Лексическая семантика. Синонимические средства языка.— М.: Наука, 1974.
3. *Апресян Ю. Д.* Лексикографические портреты (на примере глагола быть) // НТИ. Сер. 2. — 1992, № 3. — С. 20–33.
4. *Апресян Ю. Д.* Лингвистическая терминология словаря // Новый объяснительный словарь синонимов русского языка. Первый выпуск. — М.: Школа «Языки русской культуры», 1999. — С. XVI–XXXIV.
5. *Апресян Ю. Д.* Фундаментальная классификация предикатов и системная лексикография // Грамматические категории: иерархии, связи, взаимодействие. Материалы международной научной конференции. СПб., 2003. — С. 7–21.
6. *Богуславская О. Ю.* БЕЗЛЮДНЫЙ // Новый объяснительный словарь синонимов русского языка. Первый выпуск. — М.: Школа «Языки русской культуры», 1999. — С. 1–3.
7. *Васильева Н. В.* Наречие // Лингвистический энциклопедический словарь. М, 1990.

8. *Журинский А. Н.* О семантической структуре пространственных прилагательных // Семантическая структура слова. Психолингвистические исследования. — М.: Наука, 1971. — С. 96–124.
9. *Кюсева М. В., Резникова Т. И., Рыжова Д. А.* Типологическая база данных адъективной лексики // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодн. Междунар. конференции «Диалог» (Бекасово, 29 мая — 2 июня 2013 г.) Вып. 12 (19). Т. 1. М., Изд-во РГГУ, 2013. С. 367–376.
10. *Левонтина И. Б.* СЛИШКОМ // Новый объяснительный словарь синонимов русского языка. Первый выпуск. — М.: Школа «Языки русской культуры», 1999. — С. 372–377.
11. *Рахилина Е. В.* Семантика размера // Семиотика и информатика. Вып. 34. — М., 1995. — С. 58–81.
12. *Семенова С. Ю.* Наречия и предикативы в прикладном семантическом словаре // Труды Международного семинара Диалог'99 по компьютерной лингвистике и ее приложениям. Таруса, 1999. Т. 1. — С. 256–264.
13. *Семенова С. Ю.* Количественный параметр и «его» глаголы // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции Диалог'2004 («Верхневолжский», 2–7 июня 2004 г.) — М.: «Наука», 2004. — С. 536–541.
14. *Семенова С. Ю.* О таксономии актантов параметрических имен // Динамические модели : Слово. Предложение. Текст. — Сб. статей в честь Е. В. Падучевой. — М., «Языки славянской культуры», 2008. — С. 691–710.
15. *Семенова С. Ю.* Русское имя параметра: метафорические и метонимические процессы // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая — 3 июня 2012 г.). Вып. 11 (18). Т. 1. — М.: Изд-во РГГУ, 2012. — С. 568–577.
16. *Семенова С. Ю.* О спецкурсе по теоретическим и прикладным вопросам изучения русской параметрической лексики // Вестник РГГУ, 2013, № 8. — С. 228–240.
17. *Сичинава Д. В.* Наречие (2011 г.) // Русская корпусная грамматика: http://rusgram.ru
18. *Филипенко М. В.* Семантика наречий и адвербиальных выражений. — М.: «Азбуковник», 2003.
19. *Филипенко М. В.* Наречия в системе «Лексикограф» // Компьютерная лингвистика и интеллектуальные технологии. Тр. междунар. конференции Диалог'2004 («Верхневолжский, 2–7 июня 2004 г.»). — С. 650–655.

References

1. *Apresian V. Ju.* (1997), Semantics and its influence on the adverbs of effort and small degree [Semantica i ee refleksy u narechii usilija i maloi stepeni], Problems of linguistics [Voprosy jazykoznanija], no. 5, pp. 16–34.
2. *Apresian Ju. D.* (1974), Lexical semantics. Means of synonymy in language [Leksicheskaja semantica. Sinonimicheskie sredstva jazyka], Science [Nauka], Moscow.
3. *Apresian Ju. D.* (1992), Lexicographic portraits (A Case Study of the Verb byt' [to be]) [Leksikograficheskie portrety (na primere glagola 'byt')], Automatic documentation and mathematical linguistics. Issue 2 [Nauchno-tekhnicheskaia informatsiia, Serija 2], no. 3, pp. 20–33.
4. *Apresian Ju. D.* (1999), Linguistic Terminology of the Dictionary [Lingvisticheskaja terminologija slovarja], New explanatory dictionary of Russian synonyms. Issue 1 [Novyj ob'jasnitel'nyj slovar' sinonimov russkogo jazyka. Pervyj vypusk], Shkola "Jazyki russkoj kul'tury", Moscow, pp. XVI–XXXIV.
5. *Apresian Ju. D.* (2003), A Fundamental Classification of Predicates and Systematic Lexicography [Fundamental'naja klassifikatsija predikatov i sistemnaja leksikografija], Grammer Categories: Hierarchies, Links, Interaction. Proceedings of an international Conference [Grammaticheskie kategorii:ierarhii, svjazi, vzaimodeistvie. Materialy mezhdunarodnoi nauchnoi konferencii], St. Petersburg, pp. 7–21.
6. *Boguslavskaja O. Ju.* (1999), LONELY [BEZLJUDNYI], New explanatory dictionary of Russian synonyms. Issue 1 [Novyj ob'jasnitel'nyj slovar' sinonimov russkogo jazyka. Pervyj vypusk], Shkola "Jazyki russkoj kul'tury", Moscow, pp. 1–3.
7. *Vasil'eva N. V.* (1990), Adverb [Narechie], Linguistic encyclopedic dictionary [Lingvisticheskij entsyklopedicheskij slovar'], Soviet encyclopedia [Sovetskaja encyklopeija], Moscow.
8. *Zhurinskij A. N.* (1971), On the semantic structure of the dimensional adjectives [O semanticheskoj structure prostranstvennyh prilagatel'nyh], Semantic structure of a word. Studies in psycholinguistics [Semanticheskaja structura slova. Psiholingvisticheskie issledovanija], Science [Nauka], Moscow, pp. 96–124.
9. *Kyuseva M. V., Reznikova T. I., Ryzhova D. A.* (2013), A typologically oriented database of qualitative features [Tipologicheskaja baza dannyh ad'ektivnoj leksiki], Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" [Komp'juternaja linguistika i intellectual'nye tehnologii: Po materialam ezhegodnoi Mezhdunarodnoj konferencii "Dialog"], Bekasovo, no. 12 (19), vol. 1, pp. 367–376.
10. *Levontina I. B.* (1999), TOO [SLISHKOM], New explanatory dictionary of Russian synonyms. Issue 1 [Novyj ob'jasnitel'nyj slovar' sinonimov russkogo jazyka. Pervyj vypusk], Shkola "Jazyki russkoj kul'tury", Moscow, pp. 372–377.
11. *Rahilina E. V.* (1994), Semantics of measure [Semantika razmera], Semiotics & informatics [Semiotika i informatika], vol. 34, pp. 58–81.
12. *Semenova S. Ju.* (1999), Adverbs and predicatives in the NLP-aimed semantic dictionary [Narechija i predikativy v prikladnom semanticheskom slovare],

Proceedings of the International seminar Dialogue'99 on computational linguistics and its applications [Trudy Mezhdunarodnogo seminara Dialog'99 po komp'juternoj lingvistike i ee prilozhenijam], Tarusa, vol. 1, pp. 256–264.
13. *Semenova S. Ju.* (2004), Quantitative parameter and verbs [Kolichestvennyj parametr i "ego" glagoly], Computational linguistics and intellectual technologies. Proceedings of the International Conference "Dialogue'2004" [Komp'juternaja lingvistika i intellektual'nye tehnologii. Trudy Mezhdunarodnoj konferencii Dialog'2004], Verhnevolzhskij, pp. 536–541.
14. *Semenova S. Ju.* (2008), On taxonomy of the parametric nouns arguments [O taksonomii aktantov parametricheskikh imen], Dinamic models: Word. Sentence. Text. A Festschrift to E. V. Paducheva [Dinamicheskie modeli : Slovo. Predlozhenie. Tekst. Sb. Statei v chest' E. V. Paduchevoi], Languages of Slavic cultures [Jazyki slavianskikh kul'tyr], Moscow, pp. 691–710.
15. *Semenova S. Ju.* (2012), On metaphor and metonymy of the Russian parametric noun [Russkoe imja parametra: metaforicheskie i metonimicheskie protsessy], Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference "Dialogue" [Komp'juternaja linguistika i intellectual'nye tehnologii: Po materialam ezhegodnoi Mezhdunarodnoj konferencii "Dialog"], Bekasovo, no. 11 (18), Vol. 1, pp. 568–577.
16. *Semenova S. Ju.* (2013), On the special course "The Russian parametric words: theory and applications" [O spetskurse po teoreticheskim i prikladnym voprosam izuchenija parametricheskoi leksiki], Gerald of RSUH [Vestnik RGGU], no. 8, pp. 228–240.
17. *Sichinava D. V.* (2011), Adverb [Narechie], Russian corpora grammar [Russkaja corpusnaja grammatica], available at: http:// rusgram.ru.
18. *Filipenko M. V.* (2003), Meaning of adverbs and adverbials [Semantika narechij i adverbial'nyh vyrazhenij], [Azbykovnik], Moscow.
19. *Filipenko M. V.* (2004), Adverbs in the database "Lexicograph" [Narechja v sisteme "Leksikograf"], Computational linguistics and intellectual technologies. Proceedings of the International Conference "Dialogue'2004" [Komp'juternaja lingvistika i intellektual'nye tehnologii. Trudy Mezhdunarodnoj konferencii Dialog'2004], Verhnevolzhskij, pp. 650–655.

АНАЛИЗ ЛЕКСИКО-СЕМАНТИЧЕСКИХ ОСОБЕННОСТЕЙ РЕГИОНАЛЬНОЙ ПРЕССЫ (НА ПРИМЕРЕ ГАЗЕТ ГРОДНЕНСКОГО РЕГИОНА БЕЛАРУСИ)[1]

Шайкевич А. Я. (lingstat@yandex.ru),

Савчук С. О. (savsvetlana@mail.ru)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

В статье приводятся результаты применения дистрибутивно-статистического анализа к корпусу газет Гродненского региона Беларуси. Было выделено три подкорпуса — районные газеты, городская газета «Вечерний Гродно» и комментарии читателей в «Вечернем Гродно». В каждом подкорпусе с помощью дистрибутивно-статистического метода были выделены списки маркеров, которые на основе лингвистического анализа удалось сгруппировать в кластеры, отражающие как тематические, так и стилистические особенности подкорпусов.

Для районных газет ведущими оказались маркеры, связанные работой местной власти, сельским хозяйством, охраной здоровья, охраной порядка и др. Большая группа маркеров определяет стиль текстов районных газет как официальный и книжный. В «Вечернем Гродно» наряду с темами, отражающими повседневную жизнь города, неожиданно на первый план по количеству маркеров выдвинулись темы, связанные с достопримечательностями города и его историей. В стилистическом отношении газете свойственна разговорность и диалогичность. Маркеры комментариев наследуют маркеры и кластеры маркеров из подкорпуса ВГ и демонстрируют логическое завершение основных стилистических тенденций газеты. Предложенный метод может быть использован для сопоставительного анализа других корпусов текстов.

Ключевые слова: дистрибутивно-статистический анализ, корпус региональных газет

[1] Работа выполнена при поддержке РГНФ, проект № 13-24-01004 (м)

DISTRIBUTIONAL-STATISTICAL ANALYSIS OF REGIONAL PRESS (NEWSPAPERS OF GRODNO REGION OF BELARUS)

Shaikevich A. Y. (lingstat@yandex.ru),

Savchuk S. O. (savsvetlana@mail.ru)

Vinogradov Russian Language Institute
of the Russian Academy of Sciences, Moscow, Russia

The paper is an application of distributional-statistical analysis (DSA) to the sub-corpora of Grodno region newspapers corpus. The sub-corpora under study are district newspapers, "The Evening Grodno" and commentaries to the latter. With the help of DSA hundreds of keywords have been elicited for each sub-corpus. The linguistic interpretation of those three lists showed that the keywords grouped into clusters reflect both thematic and stylistic features of the sub-corpora.

The district newspapers are specific in the choice of domains (mostly of local interest) and stylistic flavor (mostly official and bookish, to some extent resembling Soviet use). "The Evening Grodno" is more colloquial stylistically; its domains are naturally connected with the day-to-day city life and some topics which were unexpected, such as a large cluster of words denoting places of interest for tourists and inhabitants of the city. The keywords of the commentaries brings the stylistic trend of "The Evening Grodno" to its logical end. The method may be used for comparative analysis of other corpora, which might bring about new results depending on the composition of the corpus.

Key words: distributional-statistical analysis, regional press, newspaper corpus, Grodno region

Введение

В статье приводятся результаты исследования лексико-семантических и стилистических особенностей корпуса газет Гродненского региона Беларуси. Исследование проводится в рамках международного проекта, выполняемого коллективами ИРЯ им. В. В. Виноградова РАН и Гродненского государственного университета им. Янки Купалы. Цель проекта состоит в том, чтобы на материале газетных текстов выявить лексико-семантические и культурно-специфические особенности русской речи на территории Гродненского региона. В качестве экспериментальной базы используется создаваемый корпус региональных газет, который составит основу нового модуля в Национальном корпусе русского языка. В настоящее время в белорусскую часть корпуса входят 5 русскоязычных газет за 2012 год: городская газета «Вечерний Гродно»

и 4 районные газеты — «Берестовицкая газета», «Островецкая правда», «Ивьевский край», «Свислочская газета», общий объем корпуса составляет около 900 тыс. словоупотреблений. Российская часть корпуса формируется из областных и районных газет, а также региональных выпусков «Комсомольской правды».

Одновременно с составлением корпуса ведется поиск наиболее эффективных методов и приемов анализа материала. Как известно, в корпусной лингвистике принято разделение на corpus-based (CBA) и corpus-driven (CDA) подходы к изучению языковых данных. Мы предполагаем использовать оба подхода. Перспективы применения первого на основе инструментария, предоставляемого НКРЯ, изложены в [Кустова, Савчук 2013]. Ко второму подходу относится дистрибутивно-статистический анализ, опыт использования которого на материале гродненского корпуса излагается в настоящей статье.

1. Метод и процедура анализа

Существуют разные меры определения неслучайности концентрации той или иной лексической единицы в конкретном подкорпусе текстов, входящем в более широкой корпус. В настоящем проекте принимается путь, намеченный в публикации [A. Shaikevich, 2001, 229–255].

Как данное принимается следующая ситуация: существует какой-то корпус текстов (и соответствующий частотный словарь), в котором выделяется подкорпус (со своим частотным словарем). Зная долю подкорпуса в общем корпусе, мы можем подсчитать математическое ожидание (m) частоты какой-то лексической единицы в подкорпусе в предположении, что вероятность появления единицы не меняется по сравнению с общим корпусом (нулевая гипотеза). Затем реальная частота единицы в подкорпусе (f) сравнивается с математическим ожиданием, и в случае значительного расхождения двух величин делается вывод о неслучайности такого расхождения. Мера неслучайности (S) определяется по формуле:

$$S = \frac{f-m-1}{\sqrt{m}}$$

Для отрицательных значений S:

$$S = \frac{f-m+1}{\sqrt{m}}$$

При S=2 следует обратить внимание на данное слово, при S=3 возникнет подозрение в неслучайности отклонения, при S=4 подозрение превращается в уверенность.

Предположим, что корпус гродненских газет включен как подкорпус в общий корпус русскоязычных белорусских и российских газет объемом около 4 млн словоупотреблений. Доля гродненского подкопуса составит 0.22, тогда степень специфичности следующих шести слов составит (F — частота в общем корпусе):

Слово	F	m	f	S
без	3802	836	842	0
больница	1067	234	236	0
белорусский	824	182	764	42
ветеранский	102	22	74	10
бизнес	1158	255	81	−11
власть	2098	462	191	−12

Вывод очевиден: частота слов *без* и *больница* ничем не отличается от общей нормы, слова *белорусский* и *ветеранский* крайне специфичны для гродненского подкорпуса, а частота слов *бизнес* и *власть* значимо меньше, чем в общем корпусе.

Будем называть лексическими маркерами те лексические единицы, реальная частота которых значимо превышает математическое ожидание (S превышает некоторый порог).[2]

В корпусе гродненских газет мы выделили три подкорпуса: районные газеты (РГ) (80% всего объема), «Вечерний Гродно» (18%) и комментарии читателей ВГ (2%). К ним применена та же процедура выявления маркеров. В подкорусе РГ обнаружено 217 маркеров (S≥2), в «Вечернем Гродно» — 1080 маркеров (S≥3), в комментариях — 360 (S≥2). Расхождения в числе маркеров в значительной мере связаны с самим устройством нашей формулы. Она легко находит маркеры в подкорпусе, составляющем небольшую долю общего корпуса (скажем, менее четверти). Когда подкорпус составляет половину общего корпуса и больше, маркеры выделяются с трудом.

Чтобы обойти эту трудность и все-таки найти характерные черты подкорпуса РГ, прибегнем к двум приемам. Первый прием возможен при данной структуре изучаемого корпуса, в котором больший и меньший подкорпусы покрывают почти весь объем корпуса. Тогда отрицательные маркеры меньшего корпуса можно использовать в качестве кандидатов в положительные маркеры большего корпуса. Рассмотрим как пример слово *ввод*. Оно встретилось в РГ 59 раз и за их пределами не встретилось ни разу. В большем подкорпусе частота 59 дает S меньше 2, но в городской газете частота 0 делает слово отрицательным маркером и тем самым кандидатом в маркеры подкорпуса РГ. Такими же кандидатами становятся *божественный*, *ветеранский*, *вклад*, *возглавлять*, *воинский*, *воспитанник*, *вправе*, *встреча*, *выборы* и многие другие слова.

Второй способ пополнения маркеров возможен при любой структуре общего корпуса. Он сводится к группировке нескольких слов, что увеличивает реальную частоту и может вести к повышению S. Слова *ВИЧ*, *ВИЧ-инфекция* и *ВИЧ-инфицированный* порознь не обладают частотой, достаточной для S=2, но вместе эти три слова набирают частоту 141 в РГ (при 145 во всем корпусе), значение

[2] Строго говоря, их следовало бы называть положительными лексическими маркерами. Отрицательными лексическими маркерами можно называть слова, чья частота существенно ниже математического ожидания, но их в корпусе обычно не так много, и они хуже интерпретируются содержательно.

S превысит 2 и сделает группу полновесным маркером.[3] Подобным образом маркерами РГ станут *выяв-ить, животновод-ство, заготов-ить, информ-ация, кредит-, назнач-ить, налог-, нарко-тик, необходим-о, поруч-ить, страхов-ой, уделять, уплат-а, Христ-ос*. Так же пополняется круг кандидатов в маркеры: *гордиться, достав-ить, предостав-ить, прокур-ор, эффективн-ость 4*. Группировку кандидатов в маркеры будем ниже давать в круглых скобках, заключая список указанием частоты в изучаемом подкорпусе на фоне остальных двух подкорпусов.

Оглядываясь постоянно на тексты, мы можем сводить маркеры в некоторые кластеры, открывающие для нас тематическую и стилистическую специфику изучаемого подкорпуса[5].

2. Анализ результатов: маркеры в районных газетах

Естественно предположить, что в районных газетах будут представлены маркеры, семантически связанные именно с данным **административным образованием**: *районный* f=1553 S=8, *район* f=2260 S=6, *сельский6 S=4, агрогородок7*, S=3 *совет* S=3 *административный*, *населенный*, *отдел*, *пункт*, *сельсовет*, *территория* (*межрайонный*, *муниципалитет*, *полномочия*, *поселковый*, *село* 226:9). К этому же кластеру следует отнести жителей района с их жильем и контактами с администрацией: *работник* f=1030 S=6, *заявление* f=431 S=3,

---

[3] Для краткости одно из слов такой группы будем делать ее представителем, выделяя в нем основу. Так за символом *выплат-а* будут скрываться четыре слова — *выплата, выплатить, выплачивать, выплачиваться*.

[4] Группировку кандидатов в маркеры будем ниже давать в круглых скобках, заключая список указанием частоты в изучаемом подкорпусе на фоне остальных двух подкорпусов.

[5] Такой анализ необходимо проводить при постоянном обращении к текстам, чтобы избежать ошибок при классификации многозначных лексем, а также контролировать влияние привычных ассоциативных связей. Так, например, первоначально предполагалось, что слово *сердце* имеет отношение к кластеру «медицина», а *минировать* — к «военной истории». Однако в ходе проверки выяснилось, что *сердце* (в районных газетах f=229 S=2) употребляется почти исключительно в метафорическом значении (*чистые сердца, доброе сердце, зов сердца, найдется место в сердце, прикипела душой и сердцем* и под.), и поэтому не годится на роль «медицинского» маркера. А слово *минировать* (в ВГ f=8, S=4) вообще, как выяснилось, к войне не имеет отношения, поскольку все 8 вхождений встретились в составе терминологического наименования *минирующая моль*.

[6] Высокий показатель маркера *сельский* связан, скорее всего, с тем, что в подкорпус РГ попали газеты четырех районов с преимущественно сельским населением.

[7] Слова *агрогородок* (в Белоруссии это официальное название одного из типов сельских населенных пунктов) и *агроусадьба* (объект экологического туризма) имеют выраженную региональную специфику, которая проявляется в различии частотных характеристик. Так, в белорусских газетах в составе регионального корпуса слово *агрогородок* встретилось 202 раза (на 0,9 млн словоупотреблений), а в российских — лишь 2 раза (на 9 млн); в корпусе современных российских СМИ (173 млн) зафиксировано всего 19 вхождений этого слова.

подача =3, *жилищ*-е, *жилой*, *житель*-, *помещение*, *проживать*, *уплат-а* (*односельчанин*, *паспорт*, *пустующий*, *сельчане* 299:15)

Но среди маркеров РГ представлены все звенья вышестоящей **власти** (кроме самого слова власть) со всеми их функциями: *организация* f=1321 S=5, *республика* f=1193 S=5, *государственный* S=3, *законодатель-ство* S=3, *исполнительный* S=3, *ответствен*- S=3, *учреждение* S=3, *документ*, *закон*, *исполнение*, *комитет*, *контроль*-, *меры*, *назнач-ить*, *пленум*, *поруч-ить*, *проверка*, *республиканский*, *управление*, *установленный* (*декрет*, *должност-ь*, *заседание*, *надзор*, *нормативный*, *президиум*, *распорядиться*, *распоряжение*, *рассмотреть*, *регламентировать*, *регулировать*, *указ* 896:47). Персонификация власти отражена в маркерах *руководитель* f=458 S=5 и *председатель* f=983 S=4 (*возглавлять*, *глава*, *управляющий* 244:17). Индивиды обозначены здесь маркерами *граждане* f=1035 S=6, *лицо* f=738 S=3, *гражданин*.

Обобщенные маркеры **«низов»** — *население* f=603 S=3, *профсоюз* f=228 S=3, *ОО* [общественное объединение], «*Белая Русь*» (*первичка* 34, *первичный* 124:3). От «верхов» исходит информация — *информ-ация*, *обращение*, *проинформировать*; информация сопровождается другими средствами убеждения, часто идеологизированными — *акция* f=319 S=3, *мероприятие* f=555 S=3 (*агитбригада*, *акцентировать*, *идеолог*-, *комсомольский*, *месячник*, *нравственн-ый*: *октябренок*, *октябрятский*, *оповещение*, *опубликование*, *пионерия*, *подчеркнуть*, *разъясн-ить*, *убеждать* 507:19). Особое значение имеет воспитание патриотизма — *ветеран*- f=364 S=4 (*афган-ский*, *беззаветный*, *воин*-, *горд-иться*, *мемориальный*, *отечество*, *партизанский*, *патриотический*, *подвиг*, *преданность*, *пьедестал*, *увековечение*, *фронтовой* 540:24).

Не только слова, но и **материальные блага** идут сверху вниз: *социальный* f=619 S=4, *обеспеч-ение* f=651 S=4, *пенси-я* S=3, *выплат-а*, *заработ-ок*, *плата*, *пособие*, *предостав-ить* (*льгот-а*, *надбавка*, *нетрудоспособн-ость*, *оклад*, *перерасчет*, *престарелый* 378:22). Благожелательство «верхов» выражено маркерами *благополучие* и *удел-ять* (*благосостояние*, *поощр-ить*, *процветание*, *удовлетворение*, *чутк-ий* 171:0).

Характерная черта советского быта, сохранившаяся в РГ, — **приуроченность** государственной благотворительности **к каким-то датам**: *праздн-ик* f=765 S=3, *наград-а* f=321 S=3, *поздрав-ление* f=332 S=3, *благодарность*, *вруч-ить*, *грамота*, *подарок*, *почетный* (*выраз-ить*, *годовщина*, *заслуг-а*, *искренн-ий*, *преддверие*, *предпраздничный*, *предстоящий*, *премия*, *приветствовать*, *признательность*, *приуроченный*, *торжественный*, *удостоить*, *чествова-ние*, *юбиляр* 1119:43).

Обратная связь снизу обеспечивается **готовностью масс к действию** (маркер *активн-ый* f=362 S=3) и участием в выборах: *выдвинуть* f=273 S=3, *депутат* f=298 S=3, *избиратель*, *кандидат* (*выборы*, *голосование*, *делегат*, *отчет-ный*, *предвыборный*, *проголосовать*, *созыв* 408:3). Главное же для «низов» — **труд на благо общества**: *работа* f=2798 S=7, *труд* f=688 S=5, *трудовой* S=4, *коллектив* (*бескорыстный*, *вдохновлять*, *дисциплина*, *добросовестный*, *передовик*, *потрудиться*, *самоотверженный*, *слаженн-о*, *созидательный*, *сплоченный*, *старательный*, *терпение*, *труженик*, *умел-ый* 429:5).

Общий итог выражен маркером **порядок** (*стабильн-ый, устойчивый* 66:2).

В районах Гродненской области, представленных в корпусе РГ, основной сектор — **сельское хозяйство**, поэтому вполне ожидаем широкий круг соответствующих маркеров: *зерно-* f=420 S=4, *сельскохозяйственный* S=3, *агропромышленный*, *гектар*, *животновод-*, *заготов-ить*, *комбайн*, *совхоз*, *участок*. Из длинного перечня кандидатов в маркеры назовем самые частые — *корм, молоко, зимний, посев, корова, скот, озимый, механизатор* (всего 80 слов, 2683:76).

Но не только сельское хозяйство, вся **экономика** образует мощный кластер маркеров в РГ. Просматривая этот список, так и слышишь оптимистические экономические отчеты на съездах КПСС: *хозяйство* f=851 S=5, *выполн-ить* f=621 S=4, *осуществ-ление* f=409 S=4, *размер* S=3, *текущий* S=3, *величина*, *достижение*, *задача*, *итоги*, *материальный*, *млн*, *ОАО, обслуживание, объединение, период, показатель, приобретение, продукция, произвести, производственный, процент, результат, рост, сроки, темп, товар, услуги, цель*. Список кандидатов включает 30 слов (например, *валовой, обязательство, отрасль, совершенство-вание, ускорение*; 1264:31). Рядом находим **финансовый** кластер, кажущийся более современным: *бюджет-* f=258 S=3, *денежный, имущество, кредит, налог, руб., страхов-ой, сумма* (*задолженность, затраты, израсходовать, наниматель, оборот, пеня, плательщик, подоходный, расходы, ревизия, рента, ссуда* 480:25).

Максимум эмоциональности сосредоточен в сверхкластере маркеров РГ, сводимом к трем словам — **семья, дети, школа**. *Семья* f=975 S=4, *брак, рождение*; *дети* f=1374 S=3, *возраст* S=3, *воспита-ние, ребенок, родитель; профессиональный* f=468 S=4, *школа* f=923 S=3, *образование* f=625 S=3, *учащийся* S=3, *филиал* S=3, *базовый*, *знание*, *культура*, *лицей*, *молодежь*, *педагог-*, *подросток*, *профессия*, *ребята*, *учитель*. Длинный список кандидатов в маркеры содержит очень разные слова: от официальных (*алименты, безнадзорный, времяпрепровождение*), высоких (*напутствие, наставник, юный*) до умиленно-ласкательных (*детишки, детвора, ребятишки*).

Сюда же примыкает и **физкультурный** кластер: *СШ* [спортивная школа] f=384 S=3, *команда*, *победитель*, *турнир*, *борьба* (*велопоход*, *гиревой*, *гребля*, *ДЮСШ, купаться, мяч, перетягивание, пионербол, подопечный, соперник, спартакиада, стритбол* и т. п., всего 26 слов 684:10). Особый кластер в РГ образует единственный маркер *ХРИСТ-ос* с очень длинным списком ассоциированных слов (*акафист, божественный, вечерня, воскреснуть, духовн-ый, молебен, Пасха, прихожане* и т. п. всего 35 слов 867:10).

Два кластера угрожают жизни гродненских районов — **болезни** и **преступники**.

Кластер «**охрана здоровья**» содержит маркеры *ВИЧ-, здоровый, здоровье, нарко-тик, профилактик-а* и множество кандидатов в маркеры (*алкоголизм, бюллетень, вирус, инфекция, клещ-, курение, очаг, сыпь* и т. п., всего 26 слов 971:64).

Неожиданным для нас оказался громадный кластер «**охраны порядка**». Казалось бы, эта тема присуща всем СМИ, однако именно районные газеты выдвигают ее на первый план: *безопасность* f=459 S=4, *несовершеннолетний* S=3, *органы* S=3, *внутренних*, *возмест-ить*, *кодекс*, *милиц-ия*, *оперативн-ый*, *правонарушение, преступ-ный, РОВД, уголовный* (более 50 кандидатов в маркеры, например *антиобщественный, арест, ДТП, кража, МВД, наказание, обвиняемый,*

опьянение, пресеч-ение, причин-ить, прокур-ор, раскрываемость, самогон, следственный, *ст*. и т. п. 2119:67).

Перечисленные выше тематические кластеры маркеров не исчерпывают специфики подкорпуса РГ. Остается большая совокупность **стилистических маркеров**, тяготеющих одновременно к официальности и книжности: *внимание* S=4, *качество* S=4, *внес-ти*, *вопрос* S=3, *проводиться*, *путем*, *являться* S=3, *данный*, *действовать*, *иметь*, *иметься*, *иной*, *категория*, *необходим-о*, *оказание*, *определ-ить, отношение, принять, провед-ение, в соответствии, соответствующий, степень, в сфере, участие, учет, в ходе*. Остается еще 36 кандидатов в маркеры (например, *аспект, вследствие, вышеназванный, значим-ый, ибо, исходя, каков, констатировать, контекст, менталитет, насущный, поприще, совокупный, спектр, столь, сторицей* 1309:65). Лишь одно слово выбивается из этого перечня — слово *районка*, 14 раз встретившееся в подкорпусе РГ.

3. Анализ результатов: маркеры в газете «Вечерний Гродно»

Анализ маркеров газеты «Вечерний Гродно» проводился аналогичным образом. Напомним, что тексты ВГ составляют не более 20% всех текстов совокупного корпуса гродненских газет, поэтому в результате применения формулы было получено более 1080 положительных маркеров с величиной отклонения S≥3. Следующим шагом было объединение маркеров в кластеры на основе семантического анализа.

Как и в подкорпусе районных газет, ожидаемым стало выделение группы маркеров, связанных городом Гродно и его жителями. Однако в отличие от районных газет город в ВГ представлен не как муниципальное образование с административным аппаратом, а как живая среда обитания, со своей инфраструктурой и организацией жизни горожан — их быта, труда и досуга.

Значительным по объему оказался кластер «**городская застройка и жилье**»: *ул-ица* S=16, *микрорайон* S=6, *проспект* 6, *переулок, планировка, застройка, квартал, спальный* и др., *здание* S=8, *дом* S=4 (*строение* S=4, *постройка*, *фасад* S=8, *двухэтажный* S=5, *трехэтажный*, *крыша* S=5, *комната* S=4, *хрущевка, подвал* S=5, *лифт* S=4, *окно* и др.). В городе выделяются **городские зоны** разного назначения: *центр* S=4 и *окраина*, *окрестности* S=4, *зоопарк* S=5, *парк* S=9, *ограждение*, *лесопарк* S=5, *рынок* S=7, *магазинчик* S=4, *салон*, *гостиниц-а* S=6. Самым многочисленным оказался субкластер «**транспорт**»: *автобус-* S=11, *троллейбус* S=4, *билет* S=10, *автовокзал*, *маршрут* S=4, *рейс* S=4, *остановка* S=4, *пересадка* S=5, *стоянка*, *такси*, *вокзал* S=8, *железнодорожный* S=5, *поезд* S=5, *вагон*, *авто*, *машина* S=10, *грузовик, мотоцикл-* S=7, *велосипед-, парковка* S=6, *ГАИ*, *поворот*, *транспорт, разметка, развязка* S=6, *светофор, трасса* S=6, *шоссе*.

«Вечерний Гродно» отличается от районных газет по двум параметрам. Первое отличие в целевой аудитории: газета ориентируется на жителей крупного города, промышленного и культурного областного центра, а читателем районных газет является житель райцентра или сельского поселения. Во-вторых, в отличие от остальных, ВГ — это газета, предназначенная для чтения

на досуге и потому в ней меньше пропагандистских материалов, и больше познавательных, просветительских и занимательных. Помимо городских новостей читателя привлекают темы искусства, досуга, в том числе путешествий, истории края и др. Все эти темы богато представлены в корпусе ВГ, что проявляется в составе маркеров.

Сильной стороной газеты ВГ является краеведческая тематика. Публикации под рубрикой «Красная книга Гродно» — очерки об истории и архитектуре Гродно — можно считать «брендом» газеты. Эти публикации — основной источник маркеров кластера «**история**»: *истор-ия* — S=8, *стар-ый* S=16, *сохраниться* S=6, *век* S=11, *столетие* S=4. Маркеры этого кластера характеризуют Гродно как город с богатым историческим наследием. Крупный субкластер связан с описанием архитектурных **исторических достопримечательностей** города: *архитектур-ный* S=4, *достопримечательность* S=4, *памятник*, *замок* S=15, *средневековый* S=5, *дворец* S=10, *крепость, баш-ня* S=9, *особняк, домик* S=7, *мельница* S=4, *колокольня* S=4, *казарма* S=8, *конюшня, мостовая, мост* S=11, *кованый* S=4, *кладбище* S=4 и др.

Еще более детально представлена **история в событиях и лицах**. Примечательно, что большинство маркеров в этой группе связано с военными действиями, победами и поражениями: *княжество* S=4, *князь* S=10, *Витовт, корол-ь* S=14, *корона* S=6, *Карл* S=9, *Сапега, Речь Посполита* S=5, *Наполеон-* S=12, *австро-венгерский, повстанцы, штрафник, шпион, царский, войска* S=5, *улан, гусар, полк* S=9, *орудие* S=6, *пушка* S=6, *война, бой* S=5, *сражение* S=6 *(атаковать, штурмовать, переправа, отступ-ать, разбит-ы), узник, гетто и др.* Среди других исторических реалий *магистрат* S=5, *купец* S=6, *Ожешко* S=11.

Кластер «**искусство и культура**», многосторонне отражает культурную жизнь города. В нем выделяются субкластеры «**театр**»: *театр-, актер, артист, режиссер, комедия, драм-а, драмтеатр, кукла, кукольный, роль, спектакль, сцена*; «**кино**» S=11, *кинотеатр* 6, *серия, фильм* 12, *экран, кадр* 4, *камера*; «**музыка, пение, танец**»: *музыкант, певица, танец, танцор, филармония, хореограф-, студия*; «**литература**»: *писатель-* 6; «**изобразительное искусство**»: *скульпт-ура* S=6 , *фотограф-ия* 7, *экспозиция, художник* 9, *вернисаж, галерея* 6, *музей* 7; кроме того *поклонник* 5, *публика, зритель* 8, *персонаж* 6, *сюжет* и др.

Продолжением этого кластера будут маркеры, объединенные более широким понятием «**досуг**»: *развлечение* S=4, *розыгрыш* S=5, *зрелище, аквапарк, бассейн, вечеринка, бар, кафе* S=8, *игра-* S=4, *клубы* (по интересам), *турист-* S=10, *путешеств-ие* S=4, *экскурсовод, яхта* S=5 и др.

Путешествия, а также приграничное положение Гродненского региона обусловливают значительную долю материалов на международные темы. Это и впечатления о поездках, и реклама туров, и отчеты о международных конкурсах и деловых встречах, и аналитические публикации о жизни соседей. Поэтому кластер «**зарубежье**» достаточно представителен: *Россия* S=4 *(российский 4, россияне)*, *Поль-ша* S=7 *(польский 11, Варшава 4), Литва* S=7 *(литовский* 6, *виленский 7, Вильнюс, Вильно), Европа* S=11 *(европейский, Евросоюз* 4*), Германия* S=6, *немецкий* 5, *Франц-ия* S=7, *Америк-а* S=9, *Итал-ия* S=8, *шведский, армянский, Лондон* S=4, *Турция* S=4, *виза* S=6.

Как представлена повседневность — работа, учеба, быт, семья, дети, забота о здоровье и охрана общественного порядка — в корпусе ВГ? Понятно, что эти темы не уникальны и занимают значительное место в других газетах. В чем специфика их разработки на страницах ВГ?

В отличие от районных газет практически отсутствует тема «**взаимодействия с властью**». В состав соответствующего кластера собраны маркеры, отражающие наиболее доступный рядовому обывателю уровень власти и формы общения с ней: *горисполком* S=11*, облисполком, мэр, анкета, бланк, заявка, запрет, очередь* и др.

Тема общественно-полезного труда также мало представлена в составе маркеров. Не велик кластер, связанный с промышленным или сельскохозяйственным «**производством**»: *завод-* S=9*, комбинат, фабрика* 5*, пивзавод* 5*, мастерская* 4*, агроусадьба* 4. Часто эти маркеры встречаются в текстах с исторической, экологической тематикой.

Зато маркеры, связанные со сферой «**предпринимательства**», представлены в большом количестве: *бизнес, бизнесмен* S=4*, бренд, деньги 7, доллар* 11*, евро* 13*, ЗАО* 5*, фирма* 5*, фирменный* 4*, инвест-иции* 8*, торги, маркетинг, менеджер, распрод-ажа* 6*, рекламный* 17*, купон 9, сертификация, скидка* 5*, стоимость* 4*, стоить* 6*, супермаркет, част-ный* 4.

Удивительно, что в отличие от районных газет, **криминальный кластер** ВГ не заполнен: единственный претендент на роль маркера — *насилие* встречается в тексте о домашнем насилии, разъясняющем и комментирующем статьи закона. Правда, как будет показано ниже, криминальная тема получает некоторое развитие в комментариях читателей ВГ.

Повседневная жизнь не может обойтись без темы «**кулинария**», чрезвычайно популярной во всех СМИ: рецепты приготовления блюд, кухонная утварь и техника, оценка качества продуктов и учреждений общепита, дегустация блюд, кулинарные конкурсы и мн. др. Этот кластер заполнен и в ВГ: *кухня, мяс-о, рыба* S=5*, кофе* 4*, картошка, спирт, сыр, торт, хлебный, лук, чеснок* и др. Сюда же можно включить и учреждения общепита: *ресторан, кафе, кофейня*.

Актуальной для горожан является тема утилизации бытовых **отходов** и связанная с ней экологическая тема: *отходы* S=*7, мусор* S=4*, контейнер, раздельный, полигон, макулатура*.

Наконец, всегда востребованы такие темы, как забота о здоровье и здоровом образе жизни, воспитание и образование детей. Кластер «**здоровье и медицина**» велик и разнообразен по составу, что свидетельствует о разных аспектах разработки этой темы в публикациях газеты: врач — пациент, профилактика — лечение, лекарства и методы лечения, медицинские учреждения и управление медициной: *больница* S=*7, клиника* 5*, поликлиника* 11*, врач* 5*, пациент* 5, *диагноз, инфаркт, курорт-* 6*, санаторий, фитнес 5, гигиена, лекарство 6, лечебный, лечиться, медучреждение, Минздрав* 5*, узи* 4*, препарат* 8*, антибиотики* 5*, целитель* и др.

Что касается образовательного кластера, то в нем преобладают маркеры, связанные с университетским образованием: *ГрГУ* S=8, *доцент*, *кафедра*, *лаборатория* 6, *ректор*, *студенческий* 4, *университет* 6, *ученый* 5.

Здесь мы затронули лишь тематические кластеры, однако серьезного внимания заслуживают группы маркеров, в которых проявляются стилистические различия между ВГ и районными газетами. В частности, огромная группа оценочной лексики (*вредный, громкий, грустный, заброшенный, знакомый, знаменитый, идеальн-ый, известный, могучий, мрачный, небольшой, невероятный, необычный, неприятный* и мн. др. — более 100 единиц) выделяется в качестве маркеров в ВГ именно потому, что в РГ оценка практически отсутствует. Это свидетельствует об относительной эмоциональной свободе и раскованности повествования, эксплицитном выражении авторской оценки. О более широком спектре и разнообразии модальности текста, помимо разъяснения, убеждения и долженствования, присущих районным газетам, свидетельствует большая группа глаголов речи (*рассказывать, говорить, объяснять, пояснять, обещать, советовать,* и др.). Группы слов, обозначающих «положение в пространстве» и «восприятие», можно рассматривать как стилистический маркер, связанный с присутствием в корпусе ВГ значительного количества текстовых фрагментов в регистре описания — пейзажей, строений, интерьеров (*внутри, возле, рядом, около, вокруг, вдоль, размещать-ся, располагаться, расположиться, стоять, виден, смотреть-ся, увидеть* и др.).

Что касается комментариев к ВГ, то они, конечно, наследуют маркеры и кластеры маркеров из подкорпуса ВГ, часто увеличивая долю глаголов. Наряду с *автобус, перекресток, пересадка, светофор* в кластере «**транспорт**» появляются *сходить, возить, ездить, поехать, пробка*. Разрастается кластер «**больница**» *врач, пациент, медицина, леч-ить, персонал, рожать, маммолог, морг, больной, лежать*. Появляется и новый кластер «**бедности**»: *зарплата, нищета, зарабатывать, бесплатный, дешев-ый, платить, продать*.

Очевидно должен появиться кластер «**переписки с газетой**»: *статья* S=22, *я* 21, *вы* 13, *написать* 12, *автор, комментарий* 9, *пожалуйста, уважа-емый* 5, *журналист, мой* 4. За ним следуют кластеры «**одобрения**» (*молодец* S=18, *спасибо* 16, *нормальный* 12, *правильно, супер, хорош-о* 8, *здорово* 6, *умница* 5, *великолепный, наконец-то* 3) и несогласия (*пиар* S=19, *ненависть, плох-о* 10, *бред, бардак, быдло, жаль, злоба, позор, тупой, урод* 7, *гадость* 5, *грустно, плевать* 4, *глупый, зря* 3). Обоснование оценки мы видим в маркерах *думать* S=13, *знать* 13, *подумать* 6, *понимать* 5, *видно, забыть, логично, умный, ум, ясно* 3.

В отсутствие цензуры в этом подкорпусе становятся маркерами и такие единицы, которые в двух других подкорпусах встречаются редко (*гей* S=14, *аборт* 12, *взятка, травести* 8, *гомофобия, отсидеть* 7, *критиковать, ориентация* 4, *забастовка* 3) или не встречаются совсем (*власть имущие, гетеро* 7).

Переход к живой речи, отмеченный в ВГ, здесь проявляется в максимальной степени как в маркерах диалога и местоименности (*бы* S=19, *там* 15, *зачем* 10, *будет, вот, нечего* 9, *если, кто туда* 8, *сколько* 7, *какой, откуда* 6 и т. п.), в общем тяготении к разговорности (*надо* S=14, *а, ага, ну, так* 12, *то, хотеть(ся)* 10, *очень* 9, *нельзя, такой, уж, прост-о* 8, *вообще, касательно* 7, *лично* 6), так и в специфических разговорных словах (*баксы* S=10, *мужик* 10, *телефончик* 7, *бюджетник* 6, *куча, товарищ* 4). Заметим, что минимальная доля данного подкорпуса в общем корпусе позволила стать маркерами словам, лишь двукратно

появившимся в тексте: *абсурд, додуматься, идиот, комменты, контрацепция, норовить, солярочка, сосать, феминистка, хамство, ханжить, хреново*.

4. Заключение и выводы

Применение метода дистрибутивно-статистического анализа на корпусе газет Гродненщины показало положительные результаты. Статистическая обработка корпусов с выделением маркеров и последующий лексико-семантический анализ соотношения маркеров выявляет существенные расхождения между содержательным, тематическим наполнением текстов и их стилистическим оформлением. Эти расхождения коррелируют с теми отличительными признаками районных и центральных газет, которые выделяются на других основаниях и на другом материале (см., например, Лысакова 1989, Федотова 2001, Кислая 2008). Кроме того, они позволяют выделять специфические особенности отдельных изданий, например, «краеведческую доминанту» в газете «Вечерний Гродно». Следовательно, метод может быть использован для проведения дальнейших исследований на корпусах региональных газет. Проверку этой гипотезы, а также выводов относительно специфики различных изданий, сделанных в ходе исследования, необходимо провести, используя другое соотношение корпусов. В дальнейшем мы планируем провести аналогичные исследования для того, чтобы: 1) попытаться выявить особенности районных российских газет на фоне региональных и городских, 2) российских и белорусских вечерних газет на фоне ежедневных; 3) сравнить российские и белорусские районные газеты между собой, 4) сравнить корпус российских региональных газет в целом и гродненский корпус.

Литература

1. *Кустова Г. И., Савчук С. О.* Изучение лексико-семантической и социокультурной специфики русской речи на территории Республики Беларусь (на материале текстов СМИ) // Труды Международной конференции «Корпусная лингвистика — 2013». Санкт-Петербург, 2013. С. 344–353
2. *Кислая Л. Н.* Редакционная политика районной прессы: на примере газет Новосибирской области. АКД. Екатеринбург, 2008
3. *Лысакова И. П.* Тип газеты и стиль публикации: Опыт социолингвистического исследования. Л.: Изд-во ЛГУ, 1989
4. *Федотова Л. Н.* Анализ содержания — социологический метод изучения средств массовой коммуникации. М.: Научный мир, 2001.
5. *Шайкевич А. Я., Андрющенко В. М., Ребецкая Н. А.* Дистрибутивно-статистический анализ языка русской прозы 1850–1870-х гг. Т. 1. М.: ЯСК, 2013
6. *Shaikevich A.* Contrastive and comparable corpora: quantitative aspects, International Journal of Corpus Linguistics, vol. 6(2), 2001, p. 229–255.

References

1. *Kustova G. I., Savchuk S. O.* (2013), The study of lexical-semantic and socio-cultural specifics of the Russian language on the territory of Belarus (on the material of mass media texts) [Izuchenie leksiko-semanticheskoj i sociokul'turnoj specifiki russkoj rechi na territorii Respubliki Belarus' (na materiale tekstov SMI)], Proceedings of the International Conference "Corpus linguistics—2013" [Trudy Mezhdunarodnoj konferencii "Korpusnaja lingvistika—2013"], St. Petersburg, pp. 344–353
2. *Kislaja L. N.* (2008), The editorial policy of the district press: the newspaper of Novosibirsk region [Redakcionnaja politika rajonnoj pressy: na primere gazet Novosibirskoj oblasti]. Abstract of PhD dissertation, Ekaterinburg
3. *Lysakova I. P.* (1989), Type of newspaper and style of publication: sociolinguistical research [Tip gazety i stil' publikacii: Opyt sociolingvisticheskogo issledovanija]. Izd. Leningradskogo universiteta, Leningrad.
4. *Shaikevich A.* (2001), Contrastive and comparable corpora: quantitative aspects, International Journal of Corpus Linguistics, vol.6(2), 2001, p.229–255.
5. *Shaikevich A. A., Andrjushhenko V. M., Rebetskaja N. A.* (2013), Distributional-statistical analysis of the language of Russian prose of 1850–1870-ies [Distributivno-statisticheskij analiz jazyka russkoj prozy 1850–1870-h gg.], Jazyki russkoj kul'tury, Moscow
6. *Fedotova, L. N.* (2001), Content analysis as the sociological method of studying mass communication [Analiz soderzhanija—sociologicheskij metod izuchenija sredstv massovoj kommunikacii]. Nauchnyj mir, Moscow.

ПЕРЛОКУТИВНЫЕ РЕЧЕВЫЕ ДЕЙСТВИЯ И ПЕРЛОКУТИВНЫЕ ГЛАГОЛЫ[1]

Шатуновский И. Б. (shatun49@mail.ru)

Международный университет природы, общества и человека «Дубна», Дубна, Россия

Ключевые слова: речевые действия, перлокутивные действия, иллокутивные акты, речевые акты, перлокутивные глаголы, совершенный вид, цель, результат

PERLOCUTIONARY SPEECH ACTIONS AND PERLOCUTIONARY VERBS

Shatunovskiy I. B. (shatun49@mail.ru)

International University for Nature, Society, and Man "Dubna", Dubna, Russia

Perlocutionary verbs like *ubezhhdat'* 'to convince / persuade', *nastaivat'* 'to insist', *ugovarivat'* ≈'to persuade', *uspokaivat'* 'to calm', *objasn'at'* 'to explain', *xvastatsy'a* 'to boast' etc. are verbs denoting perlocutionary actions. Perlocutionary actions, as defined in the paper, are unconventional actions performed by means of conventional illocutionary acts. Perlocutionary actions are aimed to achieve certain effects, goals, but they do not necessarily achieve them. Perlocutionary verbs such as *preduprezhdat'* (*to warn*), *nastaivat'* (*to insist*), *uveryat'* ('to assure') can turn into illocutionary verbs. In this case the perlocutionary text is contracted and some parts of it are taken in the meaning of the verb becoming a sign of that contraction. Perlocutionary actions and verbs can be divided into several groups according to supposed goals and effects of a perlocutionary action. They are: (1) perlocutionary actions having a clear aim which is embedded, fixed in the meaning of the verb denoting that action; this aim can be achieved or not; (2) perlocutionary actions that do not have a clear aim, but have a bundle of possible aims that are not fixed in the meanings of the corresponding perlocutionary verbs; (3) perlocutionary (and some illocutionary) actions that have a clear aim, and that aim is achieved any time the speaker does that action. These groups differ with respect to the meaning of their perfective forms. In the paper these differences are described and explanations for semantic peculiarities of the perfective forms are proposed.

Key words: speech actions, perlocutionary actions, illocutionary acts, speech acts, perlocutionary verbs, perfective, aim, effect

[1] Публикация подготовлена в рамках поддержанного РГНФ научного проекта № 14-04-00136.

1. *Перлокутивные глаголы* — это глаголы, обозначающие перлокутивные действия. Но что такое перлокутивные действия? Как отмечает М. Я. Гловинская, понятие перлокуции у Остина неясно [Гловинская 2001: 227], поэтому оно должно быть «прояснено» тем или иным образом. Исходное определение: перлокутивные действия — это неконвенциональные речевые действия более высокого (относительно иллокутивных актов) уровня, которые совершаются посредством конвенциональных иллокутивных актов (речевых актов, далее сокращенно РА). Остин об этом: «… Осуществление локутивного акта и вместе с ним иллокутивного акта может также выступать как исполнение акта другого рода. < … > Мы назовем осуществление акта этого типа осуществлением *перлокутивного* акта, или *перлокуцией*» [Остин 1986: 88]; «Иллокутивные акты конвенциональны; перлокутивные акты *не*-конвенциональны» [Остин 1986: 100]. Ярким примером перлокутивного действия является *убеждение*. Мы убеждаем кого-нибудь в истинности пропозиции P[2] посредством высказываний, которые сами по себе являются дескриптивными РА — сообщениями или констатациями. Поэтому в случае перлокутивных действий различается то, что делают, и то, чем (какими словами) это делают. Так, угрожают, сообщая: *Если вы не сделаете это, я вас убью.* В случае иллокутивных актов совпадает то, что делают, и то, чем это делают. Так, когда утверждают, что Земля плоская, то делают это словами: *(Я утверждаю, что) Земля плоская.* (Скобки показывают необязательность перформатива.)

Заметим, что перлокутивное действие не обязательно делается посредством конкретных иллокутивных актов, располагающих собственным иллокутивным глаголом, способным употребляться дескриптивно и перформативно. Более того, большинство реально употребляемых в речевой практике речевых действий (высказываний) не относятся к конкретным узким РА. Так, при совершении перлокутивных действий активно употребляются РА «широкого», недифференцированного побуждения, выражаемого формами повелительного наклонения [Рассудова 1968: 102; Падучева 1996: 79], не имеющими специального лексического показателя иллокутивной силы (перформативного слова). Аналогично, перлокутивные действия часто совершаются посредством «широких» дескриптивных (описывающих действительность) РА (высказываний), о которых можно сказать не более того, что они фактивны (среди «узких», конкретных РА к ним относятся сообщения, объявления, констатации) или нефактивны (среди «узких», конкретных РА к ним относятся утверждения, заявления, уверения, заверения) [Шатуновский 2001].

Все иллокутивные акты конвенциональны в том простом смысле, что их значение и употребление основывается на конвенциях (правилах) языка. Для того, чтобы пообещать что-то или сообщить, что идет дождь (со стороны говорящего (Г), а со стороны адресата (А) — понять, интерпретировать сказанное как обещание и сообщение соответственно), необходимо соблюдать (для А — знать) правила языка. У иллокутивных актов есть специальные языковые

---

[2] P — пропозиция, описывающая какое-то положение дел (состояние, событие, действие, процесс и т. п.) в действительности / положение дел, описываемое пропозицией.

показатели, есть конвенциональная форма, так же, как и у отдельных слов и у, скажем, синтаксических типов предложений. Это, прежде всего, перформативы и перформативные формулы (клише), формы наклонения, порядок слов, специальные частицы и т. п.

В случае перлокутивных актов (действий) не надо знать никаких правил и конвенций, ни языковых, ни внеязыковых, касающихся собственно перлокутивных актов. Всё, что здесь нужно, это знать, как совершать и интерпретировать иллокутивные акты, посредством которых они совершаются, и плюс к этому иметь экстралингвистические знания, касающиеся устройства мира и ума человека, и владеть естественной, внеязыковой, обыденной логикой, которой все люди владеют, независимо от конкретного языка. Так, высказывание *Если ты это сделаешь, я тебе голову оторву* является угрозой, но это следует из самого содержания этого сообщения, но не из наличия каких-то специальных конвенциональных показателей. Г, сообщая это А-у, обычно стремится добиться того, чтобы А не делал Р. Но если А и не станет после и в результате этих слов Г делать Р, то это не потому, что он понял, что это угроза, но исключительно потому, что ему не хочется испытать ущерб, который ему причинит Г в случае совершения им Р. Это логика жизни, а не языка.

Поскольку перлокутивные действия осуществляются посредством иллокутивных актов и в каком-то смысле «состоят» из них (как предложение строится и состоит из слов), любое перлокутивное действие состоит, по крайней мере, из одного иллокутивного акта (высказывания с определенной иллокутивной силой), но может совершаться и целым рядом, комплексом иллокутивных актов (высказываний), часто разного типа. Если перлокутивное действие состоит из одного иллокутивного акта, то в этом случае можно говорить о перлокутивном акте. Например, можно похвастаться одним высказыванием, сообщающим о каких-то положительных качествах или достижениях Г — это будет перлокутивный акт хвастовства. Однако можно хвастаться и обычно хвастаются, а тем более убеждают, уговаривают и т. д. целым рядом высказываний, объединенных одной целью. Поэтому мы употребляем здесь более общий термин *перлокутивное действие*, не предполагающий, как термин *акт*, единичность.

2. С описанной выше спецификой иллокутивных и перлокутивных актов (действий) связаны особенности употребления соотносительных с этими актами иллокутивных и перлокутивных глаголов. Иллокутивные глаголы могут употребляться и дескриптивно — для описания данного РА и отсылки к нему в другом высказывании, и перформативно — для совершения этого самого речевого акта и экспликации его характера, т. е. указания, какой именно акт совершается. Перлокутивные глаголы не употребляются перформативно [Austin 1962: 103]. Мы не убеждаем, говоря: **Я вас убеждаю, что вам надо сходить к врачу, вы плохо выглядите*, и не угрожаем, произнося: **Я вам угрожаю, что я вас убью*. Вопрос: почему? Потому, что здесь нет ничего скрытого, конвенционального, о чем сигнализировали бы слова *угрожаю*, *убеждаю* и т. п. Экспликация типа перлокутивного действия — **Я вам угрожаю…* — ничего не прибавляет к … *если вы это сделаете, я убью вас.* Перлокутивные глаголы

не употребляются (без особых оснований) в перлокутивных актах потому, что их добавление нарушает один из главных принципов речевого общения [Грайс 1985]: не говори лишнего.

3. Перлокутивные действия могут переходить в иллокутивные акты в тех случаях, когда происходит их конвенционализация. Это приводит к возникновению переходных случаев между иллокутивными и перлокутивными глаголами, когда один глагол может употребляться и перлокутивно, и иллокутивно (*предупреждаю, уверяю, умоляю, настаиваю* и т. д.). Проиллюстрируем это на примере *предупреждения*. *Предупреждение* — это перлокутивное действие, которое совершается дискурсивным «блоком», состоящим из двух естественно — с точки зрения логики речевого общения — связанных иллокутивных «частей»: сообщения, что возможно / будет некоторое плохое для A P, и побуждения принять это во внимание, учитывать в своих дальнейших действиях (≈ «X говорит это потому, что хочет, чтобы Y руководствовался этой информацией в своих действиях» [Гловинская 1993: 173]). Компонент побуждения может по-разному конкретизироваться в различных ситуациях предупреждения: ‘не делай Q, которое может привести к P’; ‘сделай Q, которое позволит предотвратить P / минимизировать его последствия’; ‘приготовься к тому, чтобы столкнуться с P или его последствиями’ и т. д.: *Не пей из лужи, Иванушка, козленочком станешь!* В подобных случаях *предупреждаю* обычно не употребляется, просто потому, что оно здесь излишне, хотя дескриптивно это действие описывается именно как предупреждение: *Из машины не вылезайте, кнопки все нажмите — папаша суровый, может и палкой заехать, и собаку спустить! — предупредил Байрам* (М. Гиголашвили. Чертово колесо)[3]. В то же время *предупреждать* может использоваться перформативно, формируя вместе с вводимой им пропозицией иллокутивный акт. В этом случае оно «вбирает» в свое значение какие-то элементы перлокутивного текста, претерпевая тем самым конвенционализацию, лексикализацию. Этот переход сигнализируется сокращением перлокутивного текста, какие-то элементы которого уже не выражаются открыто как таковые, но сигнализируются перформативным *предупреждаю*, употребление которого в этом случае становится если не необходимым, то весьма желательным: *Предупреждаю: в поле будет жарко. — Я не боюсь.* (И. Грекова. На испытаниях). В данном случае эксплицитно выражено то, что будет плохое для A P; наличие побуждения ‘прими это во внимание, подумай, может быть, тебе не стоит ехать’ сигнализируется здесь перформативом *предупреждаю*. Другой вариант — эксплицируется только побуждение, а сообщение о плохом P имплицируется перформативным глаголом *предупреждаю*: *Я просто еще раз предупреждаю тебя, чтобы ты не выходила из офиса* (Т. Устинова. Большое зло и мелкие пакости) → ‘если выйдешь, с тобой будет нечто плохое’. Ср. также замечание Серля: «Некоторые глаголы могут выступать в разных случаях с разной иллокутивной целью. … Так, …: «Предупреждаю: отстань от моей жены!» (директив), «Предупреждаю,

[3] Источником примеров в статье являются материалы «Национального корпуса русского языка».

что бык вот-вот бросится» (репрезентатив)» [Серль 1986: 194]. Как представляется, в обоих случаях это иллокутивный акт именно предупреждения (warning), просто в одном случае имплицитным остается компонент сообщения о возможном плохом для А Р (= ‘если ты не отстанешь от моей жены, я каузирую тебе плохое Р’ — разновидность предупреждения, предупреждение-угроза), во втором — компонент побуждения: ‘беги / приготовься к схватке или т. п.’.

4. Иной подход к разграничению иллокутивных и перлокутивных глаголов, представленный в российском языкознании прежде всего в работах М. Я. Гловинской [2001; 1993], опирается на критерий достижения перлокутивного эффекта (результата) («… The perlocutionary act … is the *achieving of* certain *effects* by saying something» [Austin 1962: 120], см. также [Austin 1962: 101]) и соответственно в русском языке на различие НСВ и СВ. Перлокутивы «обозначают речевой акт с реальным, достигнутым результатом, притом запланированным заранее. При этом граница между иллокутивами и перлокутивами проходит обычно внутри видовых пар: в форме НСВ (в нерезультативных значениях) глагол является иллокутивом, в форме СВ — перлокутивом» [Гловинская 2001: 278]. В соответствии с этим критерием глагол НСВ, например, *убеждать* — это иллокутивный глагол, соотносительный глагол СВ *убедить* — перлокутив. На наш взгляд, имеет смысл разграничить иллокутивные и перлокутивные глаголы и действия, не связывая это противопоставление с реальным достижением результата и различием видов. НСВ *убеждать* и др. глаголов этого типа и обозначаемые ими действия (деятельность) имеют объективные отличия от НСВ иллокутивных глаголов и соответствующих речевых действий и объективные сходства с СВ *убедить* и др. И главное — *убеждают, успокаивают* и т. д. посредством иллокутивных актов. Процесс (деятельность) убеждения так же неконвенциональна, как и достижение результата. И НСВ, и СВ в этом случае обозначают одно и то же действие, одну и ту же неконвенциональную, осуществляемую посредством конвенциональных иллокутивных актов деятельность (но с разных сторон, в разных аспектах, как это и свойственно формам вида). Поэтому если принять в качестве критерия перлокутивности достижение действием эффекта, окажется, что одно и то же речевое действие, пока оно не достигло результата, а совершается, является иллокутивным актом, а после того, как оно достигнет результата, оно же становится задним числом перлокутивным актом (действием). Более того, очень часто остается неизвестным, достигло ли результата действие, ср.: *Почему? Как? Где я вообще? — У себя дома, — успокоил его Крячко.* (Н. И. Леонов, А. В. Макеев. Гроссмейстер сыска) = ‘сказал с целью успокоить’, достиг ли результата, успокоил ли — неизвестно.

5. В то же время все перлокутивные речевые действия, поскольку это намеренные, контролируемые действия, по определению производятся с какой-то целью, для достижения какого-то эффекта. (Цель связана с эффектом, результатом действия, цель — это планируемый субъектом действия его результат, эффект, достижение которого входит в намерения Г.) Однако не всё так просто с целью и намерением — в жизни и в языке. Различные действия

и глаголы являются в разной степени и в разном смысле целенаправленными. С точки зрения наличия и характера цели перлокутивные речевые действия и обозначающие их глаголы делятся на несколько групп. Ядро, центр размытой области перлокутивных действий составляют действия и глаголы в полном смысле целенаправленные— имеющие четкую цель, фиксированную в значении слова, описывающего это действие, полностью и безусловно намеренные. Этой целью является осуществление какого-то воздействия на А, изменение состояния его «ума» — в широком смысле, ментальной сферы, включающей мысли, чувства и волю адресата. Адресат здесь является не только адресатом, но одновременно объектом воздействия. Это такие действия и глаголы, как *убеждать / убедить, уговаривать / уговорить, настаивать / настоять, успокаивать/ успокоить* и т. п. Такие действия являются полностью контролируемыми в отношении процесса совершения этого действия, выражаемого формой НСВ, но не полностью контролируемыми в отношении достижения результата, результат здесь может быть достигнут, а может быть и не достигнут. Поэтому такие глаголы могут иметь СВ со значением достижения или в отрицательной форме недостижения перлокутивной цели: *С (не) убедил А-а в* Y[4].

6. Другая группа перлокутивных глаголов / действий — это (глаголы, обозначающие) действия, не имеющие фиксированной, выделенной цели, цели, которая была бы закреплена, лексикализована в значении обозначающего это действие слова. Это такие глаголы / действия, как *жаловаться, ругать, угрожать, упрекать*, *хвалить*, *хвастаться* и многие другие. Такие действия также совершаются посредством конвенциональных действий собственно языкового уровня — иллокутивных актов. Так, когда Г *жалуется*, то он сообщает А-у о некоторых плохих для Г Р, когда он *упрекает* А-а, он, используя дескриптивные высказывания, констатирует, что он сделал для А некоторые хорошие для А Р, а А со своей стороны не сделал для Г ожидаемых на основе взаимности хороших для Г Р, когда Г *хвастается*, он сообщает А-у о некоторых положительно характеризующих Г Р, и т. д. Такие речевые действия также имеют какую-то цель или цели. Однако, в отличие от перлокутивных действий предшествующей группы, цель этих действий не фиксирована в жизни и тем самым в языке. С ними связан, ассоциируется «пучок» (только) возможных целей, нечетко отделенных друг от друга. Ср. различия в толкованиях цели перлокутивного действия упрека у разных авторов: «Основная цель упрека — указать, что упрекаемый обманул наши ожидания» [Шмелев 2002: 140] и «X говорит это, чтобы Y знал, что Р огорчает Х-а, и не делал больше Р» [Гловинская 1993: 198]. Однако возможно также, что Г старается сделать так, чтобы А чувствовал себя виноватым перед Г, испытывал угрызения совести, и в результате (всё это только возможно)

---

[4] На самом деле ситуация с достижением результата здесь более сложная: СВ некоторых глаголов этой группы всегда обозначает достижение результата, СВ других глаголов в одних случаях обозначает достижение результата, в других случаях достигнут ли результат — остается неизвестным. Это зависит от характера значения глагола — его конативности или неконативности. Однако освещение этого вопроса требует отдельной обширной статьи.

побудить его сделать что-то хорошее для Г. Помимо того, что цель этих действий более размыта, она очень часто и более «инстинктивна», эмоциональна, не рациональна, не очень ясна самому Г — а это значит, что она в меньшей степени является собственно целью. Здесь цель часто превращается в неосознаваемый или слабо осознаваемый мотив[5]. В мотиве сближаются, сливаются причина и цель. Так, человек может хвастаться не для того, чтобы почувствовать себя выше А-а (это была бы цель), но потому, что ему приятно чувствовать себя выше А-а (возможно неосознаваемый мотив). Г может *жаловаться* для того, чтобы побудить А предпринять какие-то действия для облегчения положения Г, но также просто для того, чтобы «облегчить душу», «дать выход отрицательным эмоциям» [Гловинская 1993: 196], разделить с кем-то тяжесть неприятных событий. Язык является не только рациональным орудием описания, информирования, задавания вопросов и т. д., но и средством иррационального выражения и «возбуждения» различных эмоций, сознательных, полубессознательных и бессознательных переживаний и чувств — в самом Г и А. Такое выражение и «возбуждение» в гораздо меньшей степени контролируется Г. Нет резкой границы между рациональным, намеренным и эмоциональным, непроизвольным, ненамеренным в речи (и тем самым в отражающем ее факты языке), в частности в перлокутивных действиях этой группы. Поскольку здесь неясны цели и тем более неясно, достигнуты ли они, такие глаголы имеют СВ со значением ‘сказал’ и не имеют СВ со значением достижения цели: *Г пожаловался А-у на Б* — ‘сообщил о некоторых плохих для Г действиях Б,’ чего он этим хотел достичь и достиг ли он этим чего-нибудь — неизвестно.

7. Наконец, еще одна группа перлокутивных глаголов — это глаголы, обозначающие действие, автоматически достигающее цели просто в силу произнесения соответствующих слов. Это, в частности, *предупредить* (в перлокутивном и иллокутивном употреблении) и *напомнить*. Чрезвычайно широкая цель в *предупредить* — учитывать то, что сообщил Г, в своих действиях, достигается просто в силу того, что А услышал и понял Г (а это является пресуппозитивными условиями успешности любого высказывания), он уже не может как-то это «выбросить» из ума и так или иначе будет это учитывать. Аналогично с *напомнить*. Информацию передают, = коммуникативную значимость имеют только те элементы, которые в данной коммуникативной ситуации имеют альтернативы [Яглом, Яглом 1973]. Поскольку здесь нет альтернативы достижению результата, результат достигается всегда, когда Г сказал Р, в то же время есть альтернатива в отношении речевого действия ‘сказать’ — Г мог сказать Р, а мог и не сказать Р, СВ здесь также не обозначает достижения результата, а имеет значение ‘сказал’. Точнее, здесь в коммуникативном фокусе ‘сказал’, достижение результата является импликативным компонентом [Karttunen 1971]: *Он предупредил А, что Р* — ‘Он **сказал** А-у, что Р / «Р» [и тем самым достиг результата]’. Аналогичная ситуация с рядом иллокутивных глаголов, которые

[5] О сближении в ряде случаев мотива и цели, их контаминации см. [Арутюнова 1992:15; Рахилина 1989].

имеют цель, направлены на достижение результата. Однако этот результат достигается автоматически просто в силу того, что Г правильно, с соблюдением всех условий успешности, произвел соответствующий РА. Так, если Г (с выполнением всех условий успешности этого РА) *приказал* А-у сделать Р, то А просто в силу этого о б я з а н с д е л а т ь Р. Если Г *пообещал* А-у сделать Р, он в силу этого связан обещанием, обязан выполнить Р. Поэтому СВ этих глаголов также имеет значение (= имеет в фокусе значения) ‘сказал’. Аналогичная ситуация с *сообщил* (*что Р*), которое, в отличие от *сказал*, предполагает результат — ‘А принял Р как факт в свою картину мира’. Этот результат достигается автоматически, просто в силу того, что Г сказал А-у, что Р / «Р» (при условии, что А доверяет искренности и компетентности Г [Шатуновский 2001]), поэтому достижение результата является импликацией, а в фокусе СВ ‘сказал’.

Литература

1. *Арутюнова Н. Д.* Язык цели // Логический анализ языка. Модели М.: Наука, 1992. С. 14–23.
2. *Гловинская М. Я.* Семантика глаголов речи с точки зрения теории речевых актов // Русский язык в его функционировании: Коммуникативно-прагматический аспект. М., 1993. С. 158–218.
3. *Гловинская М. Я.* Многозначность и синонимия в видо-временной системе русского глагола. М.: Русские словари; Азбуковник, 2001.
4. *Грайс Г. П.* Логика и речевое общение // Новое в зарубежной лингвистике. Вып. 16. М., 1985. С. 217–237.
5. Национальный корпус русского языка. — URL: http://www.ruscorpora.ru.— Режим доступа: свободный.
6. *Остин Дж. Л.* Слово как действие // Новое в зарубежной лингвистике. Вып. 17. М., 1986. С. 22–129.
7. *Падучева Е. В.* Семантические исследования. (Семантика времени и вида в русском языке. Семантика нарратива). М. Школа «Языки русской культуры», 1996.
8. *Рассудова О. П.* Употребление видов глагола в русском языке. М.: Изд-во Московского ун-та, 1968.
9. *Рахилина Е. В.* Отношение причины и цели в русском тексте // Вопросы языкознания. 1989. № 6. С. 46–54.
10. *Серль Дж. Р.* Классификация иллокутивных актов // Новое в зарубежной лингвистике. Вып. 17. М., 1986. С. 170–194.
11. *Шатуновский И. Б.* Дескриптивные высказывания в русском языке // Russian Linguistics. 2001. Vol. 25. N 1. P. 23–53.
12. *Шмелев А. Д.* Русский язык и внеязыковая действительность. М. Языки славянской культуры, 2002.
13. *Яглом А. М., Яглом И. М.* Вероятность и информация. М.: Наука, 1973.
14. *Austin J. L.* How to do things with words. Oxford: Clarendon Press, 1962.
15. *Karttunen L.* Implicative verbs // Language. 1971. V. 47. N 2. P. 340–358.

References

1. *Arutiunova N. D.* (1992), Language of aim [Jazyk tseli], in Logical Analysis of Language: Models of Action [Logicheskij analiz jazyka: Modeli dejstvija], Nauka, Moscow, pp. 14–23.
2. *Austin J. L.* (1986), How to do things with words [Slovo kak dejstvie], in New trends in linguistics abroad [Novoe v zarubezhnoj lingvistike], issue 17, Progress, Moscow, pp. 22–129.
3. *Austin J. L.* (1962), How to do things with words, Clarendon Press, Oxford.
4. *Glovinskaja M. Ja.* (1993), Semantics of Speech Verbs within the Framework of the Speech Act Theory [Semantika glagolov rechi s tochki zrenija teorii rechevyh aktov], in The Russian language in its functioning. Communication and pragmatics [Russkij jazyk v ego funktsionirovanii. Kommunikativno-pragmaticheskiy aspect], Nauka, Moscow, pp. 158–218.
5. *Glovinskaja M. Ja.* (2001), Polysemy and Synonymy in Tense-Aspect System of Russian verbs [Mnogoznachnost' I sinonimija v vido-vremennoj sistemerusskogo glagola], Azbukovnik; Russkie slovari, Moscow.
6. *Grice G. P.* (1985), Logic and conversation [Logika i rechevoe obshchenie], in New trends in linguistics abroad [Novoe v zarubezhnoj lingvistike], issue 16, Progress, Moscow, pp. 217–237.
7. *Jaglom A. M, Jaglom I. M.* (1973), Probability and information [Verojatnost' I informacija], Nauka, Moscow.
8. *Karttunen L.* (1971), Implicative verbs, Language, vol. 47, N 2, pp. 340–358.
9. *Paducheva E. V.* (1996), Semantic Investigations. (Semantics of Tense and aspect in Russian. Semantics of Narrative) [Semanticheskie issledovanija. (Semantika vremeni I vida v russkom jazyke. Semantika narrativa)], Shkola "Jazyki russkoj kul'tury", Moscow.
10. Russian National Corpus [Natsional'nyj korpus russkogo jazyka], available at: http://www.ruscorpora.ru.
11. *Rahilina E. V.* (1989), Cause / aim relationships in Russian text [Otnoshenie prichiny i tseli v russkom tekste], Issues in Linguistics [Voprosy jazykoznanija], N 6, pp. 46–54.
12. *Rassudova O. P.* (1968), The Use of Verbal Aspect in Russian [Upotreblenie vidov glagola v russkom jazyke], Izdatel'stvo Moskovskogo universiteta, Moscow.
13. *Searle J. R.* (1986), A classification of illocutionary acts [Klassifikatsija illokutivnyh aktov] in New trends in linguistics abroad [Novoe v zarubezhnoj lingvistike], issue 17, Progress, Moscow, pp. 170–194.
14. *Shatunovskiy I. B.* (2001), Descriptive utterances in Russian [Deskriptivnye vyskazyvanija v russkom jazyke], Russian Linguistics, vol. 25, N 1, pp. 23–53.
15. *Shmelev A. D.* (2002), Russian language and extralinguistic reality [Russkij jazyk I vnejazykovaja dejstvitel'nost'], Jazyki slavianskoj kul'tury, Moscow.

МЕТОДЫ УСТАНОВЛЕНИЯ СЕМАНТИЧЕСКИХ РОЛЕЙ ДЛЯ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

Шелманов А. О. (shelmanov@isa.ru),
Смирнов И. В. (ivs@isa.ru)

Институт системного анализа РАН, Москва, Россия

Ключевые слова: семантический анализ текста, семантико-синтаксический анализ, построение деревьев зависимости с помощью машинного обучения, семантически размеченный корпус

METHODS FOR SEMANTIC ROLE LABELING OF RUSSIAN TEXTS

Shelmanov A. O. (shelmanov@isa.ru),
Smirnov I. V. (ivs@isa.ru)

Institute for Systems Analysis of the Russian Academy of Sciences, Moscow, Russia

The paper introduces two methods for semantic role labeling of Russian texts. The first method is based on semantic dictionary that contains information about predicates, roles and syntaxeme features that correspond to the roles. It also uses heuristics and integer linear programming to find the best joint assignment of roles. The second method is data-driven semantic-syntactic parsing, which was implemented using MaltParser. It performs transition-based data-driven parsing simultaneously building a syntactic tree and assigning semantic roles. It was trained with various feature sets on SynTagRus Treebank, which was automatically enriched with semantic roles by the dictionary-based parser. We managed to automatically alleviate mistakes in the training corpus using output of the data-driven parser. We evaluated the performance of the parsers on the subcorpus of SynTagRus, which we manually annotated with semantic information. The dictionary-based parser and the data-driven semantic-syntactic parser showed close performance. Although the data-driven parser did not outperform the dictionary-based parser, we expect that it can be beneficial in some cases and has potentials for further improvement.

Keywords: semantic role labeling, semantic-syntactic analysis, data-driven dependency parsing, parser, semantic dictionary, semantically annotated corpus

1. Introduction

Semantic Role Labelling (SRL), sometimes also called shallow semantic parsing, is one of the best-known approaches to computational semantic analysis of natural languages. It supposes a simple model of text semantics, which treats a sentence or a clause as a situation (event or action) and entities represented in text as participants that play different roles in the situation. The model is widely used for information retrieval tasks, namely, information extraction, question answering, text summarization, machine translation and others.

Semantic role labelling consists of detection of sentence predicates expressing situation, identification of predicate arguments denoting participants related to the situation, and labeling the arguments with semantic roles. There is no conventional agreement about the set of semantic roles. However, there is a set of frequent roles that in some way are represented in most of the semantic theories. This set includes "AGENT"—the instigator of an action; "PATIENT"—participant that is affected by an action; "INSTRUMENT"—the mean by which an action is performed; "LOCATION"—the place of an action; "TIME"—the time of an action; "CAUSE"—cause of an action; "GOAL"—the entity, to which an action is directed, and others. The significant feature of semantic roles that distinguishes them from syntax relations is that they are extra-lingual concepts. The set of semantic roles and their granularity depend on domain and their application in solving high-level tasks.

We are developing the system that performs SRL-like semantic analysis of Russian texts. The early versions of the SRL system was already applied to solving many tasks of information retrieval [Osipov et al., 2008]. In this paper, we present two methods for semantic role labeling of Russian texts. The first one is a modification of the dictionary-based method implemented in the early versions of the system. The second method is using a data-driven transition-based parser for joint semantic-syntactic parsing of text. To train the data-driven parser we used SynTagRus—the Russian treebank [Apresjan et al., 2005], which we automatically enriched with semantic roles using the dictionary-based parser. We implemented technique to alleviate mistakes in the training corpus using output of the data-driven parser. Performance of the parsers was evaluated on the test corpus, which we manually annotated with semantic information.

The rest of the paper is structured as follows. Related work is reviewed in Section 2. Section 3 discusses the basic principles of our semantic model and two methods for semantic role labeling of Russian texts. Section 4 describes the experimental results for the described methods. Section 5 presents the analysis of errors, compares the described methods, and outlines future work.

2. Related Work

Semantic role labeling has been attracting attention of many researchers for the last 12 years. The mainstream approach to solve this task is to treat it as classification problem and to use supervised machine-learning techniques. This approach was pioneered by [Gildea and Jurafsky, 2002] and developed in many other works, which

enhanced feature set [Xue and Palmer, 2004], applied different machine learning methods [Pradhan et al., 2004], utilized inference procedures (e.g. based on integer linear programming [Punyakanok et al., 2008]), and applied other methods for global scoring (e.g. reranking [Toutanova et al., 2005]). Due to large interest in semantic parsing of natural language, several shared tasks devoted to the problem were conducted [Carreras and Marquez, 2004], [Carreras and Màrquez, 2005].

One of the tendencies in natural language processing consists in treating semantic role labeling as establishing labeled dependencies between predicate and its arguments and applying approaches developed for dependency parsing (e.g. transition-based parsing [Choi and Palmer, 2011]). We consider this technique promising for semantic-syntactic parsing of Russian and apply it in our research.

Simultaneously with SRL, ideas of combining syntactic and semantic parsing together has been developing. Reason for combining this procedures lies in a hypothesis that parsing would benefit from tight interaction between syntactic and semantic layers. Several researchers tried to exploit this idea using various techniques. [Gildea and Jurafsky, 2002] used reranking on a set of diverse syntax parses and semantic structures. [Musillo and Merlo, 2006], [Merlo and Musillo, 2008] created joint constituency based syntax parser that simultaneously with syntactic analysis assigns semantic roles to constituents. Great contribution to the development of syntactic-semantic parsing was made by participants of CoNLL shared tasks 2008 and 2009 [Surdeanu et al., 2008], [Hajic et al., 2009]. The setup of the tasks considered dependency-based approach for both syntactic and semantic parsing and encouraged participants to create parsing techniques that combine them. Six research teams in 2008 and four teams in 2009 implemented systems that combined syntactic and semantic parsing in some way. In our research we implemented the technique that is quite similar to the one proposed by [Samuelsson et al., 2008], which utilizes MaltParser [Nivre et al., 2007] for joint syntactic-semantic parsing. However, [Samuelsson et al., 2008] did not apply this technique to Russian texts and did not use an automatically generated training set.

Although SRL is considered as well-researched problem for many languages (e.g., the participants of CoNLL Shared Task 2009 successfully solved it for seven languages with the same framework) the progress of SRL for Russian is relatively slow. There are several research groups working on the problem [Anisimovich et al., 2012], [Ermakov and Pleshko, 2009], [Kuznetsov, 2012], [Smirnov et al., 2014]. However, it seems that evaluation results of SRL systems for Russian language have not been published. One of the major reasons for this is absence of semantically annotated corpora for training and objective evaluation of SRL systems for Russian. Although Framebank project [Kashkin and Ljashevskaja, 2013] addresses this issue, it is still in development and cannot be used for machine learning techniques and evaluation so far.

3. Methods for Semantic Role Labeling

This section discusses semantic model and two methods for semantic role labeling developed to process Russian texts: dictionary-based parser and data-driven syntactic-semantic parser.

3.1. Principles of Semantic Analysis of Russian Language

We use relational-situational model of text [Osipov et al., 2008], [Osipov et al., 2013] as underlining model of text semantics. The model is based on the theory of Communicative grammar of Russian language [Zolotova et al., 2004]. The core concept of the Communicative grammar is syntaxeme—minimal indivisible semantic-syntactic structures of language, which possesses atomic semantic meaning. We interpret nominal syntaxemes (expressed by heads of noun phrases or by main nouns in prepositional phrases) as participants of situations and semantic meanings of syntaxemes as semantic roles. Therefore, in our case, SRL consists of detection of predicate words, identification of nominal syntaxemes, and labeling them with meanings that depend on the predicate words.

According to the relational-situational model, meaning of a syntaxeme in Russian texts is determined by preposition, grammatical case, and categorial semantic class (CSC) of the head noun of the syntaxeme. Categorial semantic class is a generalized meaning of a word. We distinguish the following categorial classes: “concrete” (material entities), “abstract” (immaterial entities, states, processes), “personal” (person that able to act purposefully), “location”, “time”, “measure”, “measurement parameter”, and “quantity”. Two syntaxemes of identical morphological form may have different meanings if they belong to different CSCs.

We use rich inventory of universal roles (i.e. meanings) that consist of more than 80 roles. They partially coincide with basic semantic roles used in other research. The top ten most frequent roles are: “Subject”, “Object”, “Predicate”, “Locative”, “Deliberative”, “Directive”, “Causative”, “Possessive”, “Result”, “Addressee”. More complete list of semantic roles and their description are represented in [Osipov, 2011]. The Fig. 1 demonstrates an example of a sentence taken from SynTagRus and labeled with semantic roles.

Fig. 1. Example of a sentence labeled with semantic roles. “Temporative”—time or period of time of an action; “Ablative”—starting point of an action; “Directive”—ending point of a motion; “Subject”—initiator of an action; “Object”—something that is affected by an action

3.2. Semantic Role Labeling Using Semantic Dictionary

To perform semantic parsing, we use semantic dictionary, which is being developed in Institute for Systems Analysis of Russian Academy of Sciences [Zav’jalova, 2004], [Osipov et al., 2008]. The dictionary stores frames that provide information

about predicate and its semantic roles. The predicate is described by a set of predicate words with their lemmas. Information about roles includes sets of features that syntaxemes should have to obtain a specific role. These features include grammar case, preposition, and categorial semantic class. The developed semantic dictionary contains 2,856 frames and 3,585 predicate words. More detailed description of the semantic dictionary can be found in [Osipov, 2011]. Table 1 illustrates an example of a frame in the semantic dictionary for the situation expressed by the predicate word "отправить" ("send") from the example in Fig. 1. The dictionary-based semantic parser uses the semantic dictionary as the primary knowledge source for text processing.

Table 1. Frame in the semantic dictionary for the situation expressed by predicate words "отправить" ("send"), "направить" ("guide"), "послать" ("send"), "сослать" ("deport")

Semantic role	Categorial class	Preposition	Grammar Case
Ablative	Location, concrete	Из, из-за, из-под, от, с	Genitive
Addressee	Personal	К	Dative
Destinative	Any	Для	Genitive
Directive	Location, concrete	В, за, на	Accusative
Mediative	Concrete	По	Dative
Object	Any	–	Accusative
Objective	Any	За	Instrumental
		На	Accusative
Subject	Personal	–	Nominative

The input of the dictionary-based semantic parser is a list of sentences, which are split into clauses (simple sentences, participle expressions, and other locutions), tokens of each clause are organized in a dependency tree, and the clauses are linked with dependency relations. Tokens in a syntax tree have morphological features; nouns are assigned categorial semantic classes. CSCs are recognized by standalone CSC-processor, which uses additional dictionaries and heuristic rules.

The semantic parsing algorithm consists of the following major steps:

- predicate identification and search for a set of corresponding frames in the semantic dictionary;
- argument identification;
- argument classification;
- postprocessing.

The predicate identification step is performed using predicate words from the semantic dictionary and several pruning conditions that restrict found predicates to have particular part-of-speech (verb, noun, participle, etc.) and not to be modal verbs. For each found predicate, the parser searches for frames in the semantic dictionary by comparing lemma of the predicate with lemmas of predicate words in the dictionary. It is usual that more than one frame is found for a given predicate because of polysemy. Disambiguation of predicate sense and final choice of the frame are

carried out during postprocessing step. For each found predicate, argument identification, argument classification and postprocessing steps are performed independently.

The argument identification step is performed using a system of heuristic rules. Clause of the predicate is the main scope of search for its arguments. The parser runs through words in the clause, checks whether they satisfy a number of conditions, and assigns them weights (between 0 and 1). The weight of a word indicates confidence that the word is an argument of the predicate. The weight assignment is driven not by classifier or statistical measures but by the system of heuristics. These rules take into consideration features of the predicate, part of speech of the given word, syntactic links between the predicate and the word. Links between clauses also determine potential arguments. Parser takes into consideration words outside of the predicate clause, which are linked to it with clause relations. The result of this step is a set of potential arguments of the predicate—words that got non-zero confidence weights.

In the argument classification step, for each found frame and for each potential argument parser tries to assign semantic label independently from the other arguments using the semantic dictionary. It compares features of the given argument with features that correspond to a specific role recorded in the given frame in the semantic dictionary. The parser examines grammar case, categorial semantic class and preposition. Depending on features of the predicate, comparison function differently treats grammar cases of the argument and the role in the dictionary frame. For example, if a predicate is in passive voice the instrumental and nominal cases are switched before comparison. Result of this step is a set of arguments that got zero or one label for each frame.

The postprocessing step consist of choosing the best combination of labeled arguments for each dictionary frame and applying domain constraints. Constraints demand argument structure not to have duplicate labels. Arguments also differ by their confidence weights and if two arguments claim the same semantic label it is reasonable to assign label to argument with the greatest weight. The task of choosing the best combination of labeled arguments can be considered as optimization problem and can be solved by means of integer linear programming (ILP) [Punyakanok et al., 2008]. In particular, to solve this task we applied Hungarian method. When distributions of semantic roles for each frame are built, the parser finally chooses the best frame and corresponding distribution of roles as the result. The best frame of the predicate is determined as the frame, which leads to the greatest number of assigned semantic roles in text.

The output of the parser consists of arguments labeled with semantic roles and predicate words labeled with descriptors of the best frames.

3.3. Semantic Role Labeling Using Data-Driven Semantic-Syntactic Parsing

Semantic role labeling can be performed using dependency parser that has ability to label relations between words. The labels of relations can be considered as semantic roles. Therefore, such dependency parser can perform joint semantic-syntactic analysis simultaneously building the syntax tree and labeling tokens with semantic roles. Although semantic relations do not always coincide with syntax relations and

parser cannot determine predicates, using such semantic structure and external predicate labeler it is possible to restore semantic dependencies quite well.

To perform syntactic-semantic analysis we used MaltParser [Nivre et al., 2007]—the system for data-driven dependency parsing. MaltParser implements transition-based parsing framework, which includes various parsing algorithms and classifiers for transition prediction. It can build a dependency tree and label dependencies simultaneously. The framework provides ability to create complex features for classifiers including features that are based on partially built syntax tree and labels of established dependencies. Thus, MaltParser can perform joint semantic-syntactic analysis since it consults semantic labels and syntactic features during inference procedure.

To perform semantic-syntactic analysis MaltParser have to be trained on syntactically and semantically annotated corpus. There is only one substantial Treebank of Russian texts—SynTagRus corpus [Apresjan et al., 2005], which is part of National Corpus of Russian Language[1]. The corpus in conjunction with MaltParser was used for data-driven syntactic parsing of Russian texts by several researchers [Nivre et al., 2008], [Sharoff and Nivre, 2011]. However, there is still no substantial semantically annotated corpus for Russian for effective semantic role labeling using machine learning like PropBank [Palmer et al., 2005].

To overcome this problem we automatically annotated SynTagRus using our dictionary-based semantic parser described in 3.2. The dictionary-based parser was fed with gold syntax trees, gold morphology features, but automatically generated categorial semantic classes of nouns. The created corpus was used to train MaltParser to perform joint semantic-syntactic analysis with various features sets.

4. Experiments

This section describes experiments with the dictionary-based semantic parser and the data-driven semantic-syntactic parser.

4.1. Test Corpus and Evaluation Metrics

To evaluate performance of the semantic parsers we manually annotated subcorpus of SynTagRus with semantic information. The created semantic test corpus was annotated with nominal syntaxemes, their semantic roles and categorial semantic classes, predicates, and semantic dependencies between predicates and syntaxemes. However, not every nominal syntaxeme was annotated. The corpus contains annotations only for cases that can be found in the semantic dictionary. This limitation was introduced due to complexity of the task of manual semantic annotation of texts with rich set of semantic roles. It is also worth noting that the created corpus is not error-free and work on it is still in progress. The test corpus currently contains 1,730 sentences (29,041 tokens without punctuation), 3,871 tokens have semantic roles, and 61 roles are unique.

[1] Available at http://www.ruscorpora.ru

To evaluate performance of the semantic parsers we used three measures: precision, recall and F_1-measure. Evaluation took into account only tokens that have semantic roles in the test corpus. In this case, recall is the percentage of tokens that were properly assigned semantic role by a parser among all tokens with role labels in the test corpus; precision is the percentage of tokens that were properly assigned semantic role among all tokens that were assigned semantic role by a parser and have role labels in the test corpus. F_1-measure is the harmonic mean of precision and recall.

4.2. Experiments with the Dictionary-Based Semantic Parser

Before testing the dictionary-based semantic parser, we evaluated performance of the CSC-processor on the created test corpus. Since categorial semantic classes in the test corpus are assigned only to tokens that possess semantic role, we measured only precision of the processor. Precision is the percentage of tokens assigned proper CSC among all tokens that have categorial semantic class in the test corpus. The precision of the CSC-processor is 88.3%.

We evaluated performance of the dictionary-based semantic parser for the following cases:

- GoldSynt+GoldCSC—the input of the semantic parser consists of gold syntax trees and gold CSCs from the test corpus.
- GoldSynt+CSC—the input of the semantic parser consists of gold syntax trees from the test corpus and CSCs that are automatically generated by the CSC-processor. This case was used for generating the training corpus for the data-driven semantic-syntactic parser.
- Synt+GoldCSC—the input of the semantic parser consists of syntax trees that are automatically generated by MaltParser and gold CSCs from the test corpus. MaltParser for syntactic parsing was trained on the 48,096 sentences (74,665 tokens without punctuation) of the SynTagRus treebank with P3 feature set (see subsection 4.3). The unlabeled attachment score of the syntax parser is 88.0%.
- Synt+CSC—the input of the semantic parser consists of syntax trees and CSCs that are both generated automatically.

In all cases, morphological and lexical features were gold and were taken from the test corpus. To split sentences into clauses we used freely available NLP framework AOT.RU [Sokirko, 2001]. Table 2 shows performance of the dictionary-based semantic parser.

Table 2. Performance of the dictionary-based semantic parser

Case	Recall,%	Precision,%	F_1-measure,%
GoldSynt + GoldCSC	**82.5**	**94.5**	**88.1**
GoldSynt + CSC	70.4	89.3	78.7
Synt + GoldCSC	78.7	94.0	85.7
Synt + CSC	67.3	88.7	76.5

4.3. Experiments with the Data-Driven Semantic-Syntactic Parser

We trained the data-driven semantic-syntactic parser with four different feature sets. The basic feature set was used in [Sharoff and Nivre, 2011] to train syntax parser for Russian[2]. The basic set contains information about word form (FORM), lemma (LEMMA), part-of-speech (POSTAG), morphological attributes (FEATS), types of built relations (DEPREL). All of these features (except for DEPREL) were gold during training and testing. The tested feature sets were the following:

- P1 = Basic = FORM + LEMMA + POSTAG + FEATS + DEPREL;
- P2 = Basic + CSC (automatically generated categorial semantic classes of nouns);
- P3 = P2 + SPLFEATS (in this feature set we treat case, gender, and number as the separate features);
- P4 = P3 + descriptors of predicate frames identified by the dictionary-based semantic parser.

We trained MaltParser with LIBLINEAR library (optimized implementation of SVM without kernel function) [Fan et al., 2008] with "nivreeager" parsing algorithm [Nivre et al., 2007]. The training corpus contains 48,096 sentences (699,708 tokens without punctuation). Table 3 shows performance of the data-driven semantic-syntactic parser.

Table 3. Performance of the data-driven semantic-syntactic parser

Feature set	Recall,%	Precision,%	F_1-measure,%
P1	59.9	86.3	70.7
P2	60.4	86.5	71.1
P3	60.7	86.7	71.4
P4	**64.9**	**87.3**	**74.5**

Since the dictionary-based parser does not work perfectly, the training corpus contains mistakes. Error rate can be estimated by performance of the dictionary-based parser used to prepare training corpus that corresponds to the case "GoldSynt + CSC" in table 2.

To increase performance we suggested automatically enhancing the training corpus. We parsed the training corpus by the semantic-syntactic parser that had been trained with P4 feature set. Then we complemented training corpus with semantic labels that were found in the output of the semantic-syntactic parser but were absent in the initial training corpus. In addition, we removed sentences, which semantic labels seriously diverged from the output of the semantic-syntactic parser. From training corpus, we removed sentences, for which half and more semantic roles diverge from roles generated by the data-driven parser. Divergence was counted only if token was labeled with role both in the output of the parser and in the training corpus. We found that percentage of such sentences is less than 1%. Removing them helps

[2] Available at http://corpus.leeds.ac.uk/mocky/

to eliminate most noisy training examples that appear due to imperfect results of the dictionary-based parser. The enhanced corpus was used for training of new semantic-syntactic parser with the same feature set. We performed three iterations of this procedure. Table 4 shows performance of the semantic-syntactic parser on each iteration.

Table 4. Performance of the parser after enhancing the training corpus

#Iteration	Recall, %	Precision, %	F_1-measure, %
1	65.8	87.5	75.1
2	**66.3**	**87.7**	**75.5**
3	66.1	87.4	75.3

5. Discussion

The dictionary-based semantic parser suffers from mistakes in clause boundary recognition module. Performance of the parser also strongly correlates with quality of categorial semantic class recognition because it uses strict comparison between features of syntaxemes in text and features of roles in the semantic dictionary. However, the parser is less sensitive to errors in syntactic parsing because syntax relations only influence argument identification but not the argument labeling procedure. Evaluation of the dictionary-based parser also revealed mistakes in the semantic dictionary and some conflict situations while choosing proper semantic frame for a predicate.

Performance of the data-driven semantic-syntactic parser depends on performance of the dictionary-based parser and their mistakes notably correlate. The procedure for training corpus enhancement appeared to be beneficial. Using this procedure, we substantially increased performance of the semantic-syntactic parser. After two iterations of the procedure, both recall and precision of the parser increased. Recall increased by 1.3%, precision—by 0.4%, and F_1-measure—by 1.0%. The third iteration of the procedure decreased performance, which can be result of overfitting.

Performance comparison of the data-driven and dictionary-based parses should be done using the case “Synt + CSC” from table 2. Comparison of the best result for the data-driven parser from table 4 with the result of the dictionary-based parser in this case shows that performance of the data-driven parser is slightly lower than performance of the dictionary-based parser (the difference in F_1-measure is 1.3%). However, the used evaluation framework cannot provide purely fair comparison because the test set is not full and covers only cases that are represented in the semantic dictionary. Slightly more than 30% of roles found in the output of the parsers are not covered in the test set and could not be evaluated. The plausible case, in which data-driven parser could outperform the dictionary-based parser, is the case of unknown predicates. To test the ability of the data-driven parser to produce semantic labels for unknown predicates we plan to conduct additional experiments. For example, we could remove some predicates that appear in the test set from the semantic

dictionary making these predicates "unknown", then perform the described experiment and examine the cases with "unknown" predicates in the test set.

We intend to prepare semantically annotated test corpus with finer grained set of semantic roles (from 10 to 20 roles) and evaluate parser on it. The finer grained set would reduce labor of annotators, the amount of mistakes, and would make task of complete semantic annotation of the corpus feasible. We are also planning to include some additional features into the feature set and implement two data-driven parsers to experiment with techniques of co-training.

Good technique for evaluation of natural language components is performance evaluation of the final NLP application. Therefore, we intend to test the created SRL components as a part of IR systems that solve high-level tasks: question answering and information extraction from medical texts for English and Russian languages.

Acknowledgments

We are grateful to Laboratory of Computational Linguistics of Information Transmission Problems of Russian Academy of Sciences for provision of SynTagRus corpus. This work was partially supported by RFBR, project 13-04-12062.

References

1. *Anisimovich K. V., Druzhkin K. J., Minlos F. R., Petrova M. A., Selegey V. P. and 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 2012", Vol. 2, pp. 91–103
2. *Apresjan J. D., Boguslavskij I. M., Iomdin B. L., Iomdin L. L., Sannikov A. V., Sannikov V. G. and Sizov L. L.* (2005), Syntactically and semantically annotated corpus of Russian language: Present state and perspectives [Sintaksicheski i semanticheski annotirovannyj korpus russkogo jazyka: sovremennoe sostojanie i perspektivy], National Corpus of Russian Language: 2003–2005 [Natsional'nyj korpus russkogo jazyka: 2003–2005], pp. 193–214, (In Russian)
3. *Carreras X. and Màrquez L.* (2004), Introduction to the CoNLL-2004 shared task: Semantic role labeling, Proceedings of CoNLL-2004 Shared Task
4. *Carreras X. and Màrquez L.* (2005), Introduction to the CoNLL-2005 shared task: Semantic role labeling, Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005), pp. 152–164
5. *Choi J. D. and Palmer M.* (2011), Transition-based semantic role labeling using predicate argument clustering, Proceedings of the ACL 2011 Workshop on Relational Models of Semantics, pp. 37–45
6. *Ermakov A. E. and Pleshko V. V.* (2009), Semantic interpretation in text processing computer systems [Semanticheskaja interpretatsija v sistemah komp'juternogo analiza teksta], Information Technologies [Informatsionnye tehnologii], Vol. 6, pp. 2–7, (In Russian)

7. *Fan R.-E., Chang K.-W., Hsieh C.-J., Wang X.-R. and Lin C.-J.* (2008), LIBLINEAR: A library for large linear classification, The Journal of Machine Learning Research, Vol. 9, pp. 1871–1874
8. *Gildea D. and Jurafsky D.* (2002), Automatic labeling of semantic roles, Computational Linguistics, Vol. 28, pp. 245–288
9. *Hajic J., Ciaramita M., Johansson R., Kawahara D., Mart M. A., Màrquez L., Meyers A., Nivre J., Padó S., Štepánek J. et al.* (2009), The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages, Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task, pp. 1–18
10. *Kashkin E. V. and Ljashevskaja O. N.* (2013), Semantic roles and construction net in Russian FrameBank [Semanticheskie roli i set' konstruktsij v sisteme FrameBank], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2013" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2013"], Vol. 1, pp. 325–343, (In Russian)
11. *Kuznetsov I. O.* (2012), Automatic identification of arguments of verbs: Theoretical background and state-of-the-art techniques [Avtomaticheskoe vydelenie glagol'nyh aktantov: teoreticheskaja osnova i aktual'nye podhody], Sci-tech information. Series 2: Information Processes and Systems [Nauchno-tehnicheskaja informatsija. Serija 2: Informatsionnye protsessy i sistemy], (12), pp. 2–7, (In Russian)
12. *Merlo P. and Musillo G.* (2008), Semantic parsing for high-precision semantic role labelling, Proceedings of the Twelfth Conference on Computational Natural Language Learning, pp. 1–8
13. *Musillo G. and Merlo P.* (2006), Accurate parsing of the proposition bank, Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, pp. 101–104
14. *Nivre J., Boguslavsky I. M. and Iomdin L. L.* (2008), Parsing the SynTagRus Treebank of Russian, Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pp. 641–648
15. *Nivre J., Hall J., Nilsson J., Chanev A., Eryigit G., Kübler S., Marinov S. and Marsi E.* (2007), MaltParser: A language-independent system for data-driven dependency parsing, Natural Language Engineering, Vol. 13, pp. 95–135
16. *Osipov G.* (2011), Methods of artificial intelligence [Metody iskusstvennogo intellekta], FIZMATLIT, Moscow, (in Russian)
17. *Osipov, G., Smirnov, I. and Tikhomirov, I.* (2008), Relational–situational method for search and analysis of texts and its applications [Reljatsionno-situatsionnyj metod poiska i analiza tekstov i ego prilozhenija], Artificial Intelligence and Decision Making [Iskusstvennyj intellekt i prinjatie reshenij], (1), pp. 3–10 (in Russian)
18. *Osipov G., Smirnov I., Tikhomirov I. and Shelmanov A.* (2013), Relational–Situational Method for Intelligent Search and Analysis of Scientific Publications, Proceedings of the Workshop on Integrating IR technologies for Professional Search, in conjunction with the 35th European Conference on Information Retrieval (ECIR'13), Vol. 968, CEUR Workshop Proceedings

19. *Osipov G., Smirnov I., Tikhomirov I. and Zavjalova O.* (2008), Application of Linguistic Knowledge to Search Precision Improvement, Proceedings of 4th International IEEE conference on Intelligent Systems, Vol. 2, pp. 17-2–17-5
20. *Palmer, M., Gildea, D. and Kingsbury, P.* (2005), The proposition bank: An annotated corpus of semantic roles, Computational Linguistics, Vol. 31, pp. 71–106
21. *Pradhan S. S., Ward W., Hacioglu K., Martin J. H. and Jurafsky D.* (2004), Shallow semantic parsing using support vector machines, Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, pp. 233–240
22. *Punyakanok V., Roth D. and Yih W.-t.* (2008), The importance of syntactic parsing and inference in semantic role labeling, Computational Linguistics, Vol. 34, pp. 257–287
23. *Samuelsson Y., Täckström O., Velupillai S., Eklund J., Fišel M. and Saers M.* (2008), Mixing and blending syntactic and semantic dependencies, Proceedings of the Twelfth Conference on Computational Natural Language Learning, pp. 248–252
24. *Sharoff S. and 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", pp. 591–604
25. *Smirnov I. V., Shelmanov A. O., Kuznetsova E. S. and Hramoin I. V.* (2014), Semantic-syntactic analysis of natural languages. Part II. Method for semantic-syntactic analysis of texts [Semantiko-sintaksicheskij analiz estestvennyh jazykov Chast' II. Metod semantiko-sintaksicheskogo analiza tekstov], Artificial Intelligence and Decision Making [Iskusstvennyj intellekt i prinjatie reshenij], (1), pp. 95–108 (in Russian)
26. *Sokirko A.* (2001), A short description of Dialing Project, available at: http://www.aot.ru/
27. *Surdeanu M., Johansson R., Meyers A., Màrquez L. and Nivre J.* (2008), The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies, Proceedings of the Twelfth Conference on Computational Natural Language Learning, pp. 159–177
28. *Toutanova K., Haghighi A. and Manning C. D.* (2005), Joint learning improves semantic role labeling, Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 589–596
29. *Xue N. and Palmer M.* (2004), Calibrating features for semantic role labeling, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 88–94
30. *Zav'jalova, O. S.* (2004), About principals of creating dictionary of verbs for automatic text processing [O printsipah postroenija slovarja glagolov dlja zadach avtomaticheskogo analiza teksta], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2004" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2004"], (In Russian)
31. *Zolotova, G. A., Onipenko, N. K. and Sidorova, M. J.* (2004), Communicative grammar of Russian language [Kommunikativnaja grammatika russkogo jazyka], Institute of Russian language named after V. V. Vinogradov [Institut russkogo jazyka RAN im. V. V. Vinogradova], (in Russian).

КОРПУС ДИАЛЕКТНЫХ ТЕСТОВ В НАЦИОНАЛЬНОМ КОРПУСЕ РУССКОГО ЯЗЫКА: СЕГОДНЯШНЕЕ СОСТОЯНИЕ И ПЕРСПЕКТИВЫ[1]

Сичинава Д. В. (mitrius@gmail.com)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Качинская И. Б. (kacza@yandex.ru)

Московский государственный университет им. М. В. Ломоносова, Москва, Россия

Диалектные тексты, особенно в транскрипционной записи, часто оказываются малодоступными даже для специалистов. Уже само название Диалектного подкорпуса НКРЯ — «Корпус диалектных текстов» — указывает на то, что Пользователю предоставляется возможность работать с цельными текстами, записанными в полевых условиях в разное время. В докладе рассказывается о подготовке текстов к размещению на сайте НКРЯ в программе «Рабочее место диалектолога», с помощью которой осуществляется и разметка на уровне грамматики, в том числе с указанием диалектных особенностей, и метаразметка: указывается «паспорт» текста (место, время, автор записи, сведения об информантах и проч.), отмечаются его фонетические особенности, жанровая и тематическая отнесенность.

Ключевые слова: Корпусная лингвистика. Русская диалектология. Национальный корпус русского языка. Диалектный подкорпус

[1] Работа над Диалектным подкорпусом НКРЯ поддержана грантом РГНФ № 14-04-12012, проект «Корпус диалектных текстов Национального корпуса русского языка: пополнение и разметка», рук. Д. В. Сичинава.

THE DIALECTAL SUBCORPUS WITHIN THE RUSSIAN NATIONAL CORPUS: TODAY AND TOMORROW

Sitchinava D. V. (mitrius@gmail.com)
Vinogradov Russian Language Institute
of the Russian Academy of Sciences, Moscow, Russia

Kachinskaya I. B. (kacza@yandex.ru)
Lomonosov Moscow State University, Moscow, Russia

The main results of the project aimed at developing the dialectal subcorpus of the RNC were the creation of a pilot corpus and the change of the markup principles encompassing many dialectoloigical parameters. A working place program was created and many texts were marked up using the new technology. The present goal of our team is a considerable increase of the corpus, its representativeness and the depth of linguistic processing. The dialectal texts available for search in the RNC (www.ruscorpora.ru/search-dialect.html) will be considerably updated, with the overall corpus size reaching 1 mln tokens. The texts, mainly unpublished or published in rather obscure editions, are to be made available for a wider circle of dialectologists. Some texts are to be accompanied with video and audio. Alongside with word-by-word grammarical markup with resolved homonymy, the texts are to be tagged extensively on the metalevel (data of creation, dialect, overall phonetical properties and others). The accumulation of dialectal texts will be continued, the dialectologists who had collected valuable texts are invited to share their results with the professional community.

Kew words: Russian dialectology, corpus linguistics, Russian National Corpus

1. Существующие наработки в области диалектных корпусов

«Корпус диалектных текстов» входит в состав Национального Корпуса Русского языка (НКРЯ). Этот корпус сопоставим с такими известными национальными корпусами, как Британский и Чешский. За рубежом существуют и корпуса, включающие диалектные тексты — например, корпус, созданный в Китае (в рамках «Программы 863») или в странах Скандинавии (корпус http://www.tekstlab.uio.no/nota/scandiasyn/); много занимаются изучением народных говоров в Польше (http://www.dialektologia.uw.edu.pl/index.php); недавно появился корпус грузинских диалектов (http://mygeorgia.ge/gdc/).

В России интерес к прикладной лингвистике, созданию словарей, в том числе диалектных, сопровождается также и большой работой по созданию Диалектных корпусов, которые далеко не всегда имеют локальную

ограниченность. Так, создан сайт «Школьный диалектологический атлас "Язык русской деревни"»: http://gramota.ru/book/village. Имеются значительные по объему корпуса диалектных текстов в Казани: «Электронная библиотека русских народных говоров», Казанский (Приволжский) федеральный ун-т, где собраны материалы многих экспедиций в различных говорах Европейской части России (http://dialekt.rx5.ru/index.html); в Ижевске — Лингвогеографическая система «Диалект», Удмуртский ун-т (http://lgw2.udsu.ru:9001/). В интернете появились тексты из Шатурского р-на Московской обл. и Харовского р-на Вологодской обл. в рамках проекта «Электронные базы данных по русским народным говорам» (авторы — С. А. Крылов и А. В. Тер-Аванесова — http://starling.rinet.ru/cgi-bin/main.cgi?root=ruscorpora&encoding=utf-rus). Открылся новый сайт «Региональная этнолингвистика» с материалами по кубанским говорам (http://www.ethnolex.ru/). Во многих вузах продолжается работа по созданию и совершенствованию корпусов (пока без доступа в Интернете). По материалам трех русских говоров (двух южных и одного северного) созданы Диалектные корпуса в Центре изучения народно-речевой культуры Саратовского государственного университета им. Н. Г. Чернышевского (руководители — проф. В. Е. Гольдин и проф. О. Ю. Крючкова). Материал более чем из ста говоров Архангельской области содержится в корпусе «Электронная картотека „Архангельского областного словаря“» (МГУ имени М. В. Ломоносова), общий общем которого приближается к 2-м млн «карточек» — как видно из названия, этот корпус имеет жесткую лексикографическую направленность. Вышло несколько выпусков «Тамбовской фонохрестоматии» (Тамбовский университет), в которой расшифрованные тексты даны в сопровождении аудиоматериалов, имеется карта области, разделенная на районы, включена система Поиска, т. е. по сути эта фонохрестоматия является корпусом. Ведется активная работа по созданию корпусов по русским народным говорам в Томске, Тюмени, Челябинске, Смоленске и других научных центрах, организованных на базе университетов (см., напр., [Русская устная речь, 2011; Юрина, 2011]).

Во всех этих корпусах по-разному решаются возникающие перед диалектологами проблемы отражения фонетики, грамматики, лексики, часто они жестко направлены на исследования, традиционно проводимые лингвистическими кафедрами соответствующих вузов.

2. Концепция корпуса

«Корпус диалектных текстов» НКРЯ предполагает включение **любых** диалектных текстов на русском языке, записанных как на территории исконного проживания русского населения (Европейская часть России), так и на территориях раннего заселения (Русский Север), позднего заселения (Сибирь, Дальний Восток, Дон, Нижнее Поволжье) и миграций (говоры старообрядцев Латгалии, Азербайджана, Румынии, Австралии, Канады, Америки). Туда войдут полевые записи, аудио- и видеорасшифровки, тексты из хрестоматий, малодоступных и малотиражных сборников и изданий. Мы надеемся, что со временем этот

корпус станет репрезентативным собранием диалектных текстов и будет одним из самых посещаемых и востребованных пользователями.

Работа над Диалектным корпусом уже была поддержана грантами РГНФ: в 2006–2008 гг. (№ 06-04-13818в: «Создание корпуса диалектных и фольклорных текстов на русском языке», рук. В. М. Живов) и в 2009–2010 (№ 09-04-12159в: «Корпус диалектных текстов Национального корпуса русского языка: Грамматическая, фонетическая и метатекстовая разметка. Новый стандарт подачи», рук. В. М. Живов). Результатом первого (пилотного) проекта было создание Диалектного подкорпуса в составе НКРЯ (см. [Летучий 2005; 2008]), результатом второго проекта стала разработка нового стандарта подачи текстов и их обработки, благодаря чему появился новый системный продукт «Рабочее Место диалектолога», в котором осуществляется разметка диалектных текстов на всех уровнях: метатекстовом и грамматическом; оказалось возможным представить текст с ударениями в фонетической транскрипции двух видов: «начальной» и «облегченной», унифицированной; была значительно усовершенствована грамматическая разметка; пополнился банк диалектных текстов; часть текстов публикуется в новом формате [Качинская, 2009; 2011].

Программа, в которой непосредственно производится разметка и метаразметка диалектных текстов, — среда «Рабочее место диалектолога» (автор Т. А. Архангельский), — находится в свободном доступе.

Диалектные тексты, особенно в транскрипционной записи, часто оказываются малодоступными даже для специалистов. В ближайшее время наша задача состоит в резком увеличении количества текстов, включенных в Диалектный корпус НКРЯ (www.ruscorpora.ru/search-dialect.html) и репрезентативности географии записи. Фундаментальность проекта связана не только с количеством текстов, но и со степенью их лингвистической обработки: все тексты будут представлены с так наз. «снятой омонимией»: разметка осуществляется как на уровне грамматики (в том числе с указанием диалектных особенностей на уровне слова), так и на уровне метаразметки с указанием особенностей жанра, тематики и фонетических особенностей (в самых общих чертах). Лингвисты и все любители народного слова получат широкий доступ к цельным диалектным текстам: некоторые из них хотя и публиковались типографским способом, но обычно малыми тиражами и в малодоступных изданиях; многие тексты ранее нигде не публиковались. Некоторые тексты, записанные в последние годы на цифровые носители, предполагается сопровождать аудио- и видеоматериалами.

3. Пополнение и развитие корпуса на текущем этапе

Создание диалектных корпусов — дело во многом новое. Сбор диалектных текстов для включения их в корпус на сайте НКРЯ не должен препятствовать созданию диалектных корпусов, так сказать, «местного масштаба». С одной стороны, мы хотели бы задать некоторый стандарт подачи диалектного текста — речь идет прежде всего о текстах, которые будут специально расшифровываться для корпуса. С другой стороны, нельзя игнорировать уже имеющиеся

образцы записи народных говоров, которые оказались далеки от нашего «стандарта». Главное отличие Диалектного корпуса НКРЯ от других русских диалектных корпусов видится нам в установке на **сплошную грамматическую разметку** текстов, что соответствует общей стратегии всего Национального корпуса в целом, тогда как региональные корпуса, скорее всего, в основу разметки будут класть приципы семантики.

Работа по пополнению корпуса является достаточно сложной. Уже с самого начала эту работу приходится разделить на 2 составляющие. Первая часть — **создание банка диалектных текстов**. Тексты предоставляются диалектологами, ведущими полевую работу. Это следующие типы текстов:

1) Записи из полевых тетрадей или аудиорасшифровок, уже введенные в компьютер.
2) Опубликованные тексты, предоставленные в компьютерном варианте.
3) Тексты, опубликованные типографским способом и до сих пор **не** введенные в компьютер: это образцы говоров из хрестоматий по русской диалектологии, из учебников и пособий по русской диалектологии, пособий по изучению региональной лексики, различных сборников и статей по русской диалектологии. Все эти материалы требуется первоначально ввести в компьютер, и время на их ручной ввод или исправление текстов, полученных путем автоматического распознавания, примерно одинаковое, т. к. диалектные тексты, как правило, подаются в транскрипции с использование диакритик.

Мы также надеемся, что для Диалектного корпуса НКРЯ будут специально делаться также

4) расшифровки аудио- и видеофайлов в заданном нами стандарте фонетической транскрипции (на уровне Текста-1 или Текста-2) в шрифтах юникода.

Вторая часть работы, связанная с лингвистической обработкой текстов, тоже достаточно трудоемка.

1) Сначала требуется **подготовить** текст для его обработки в среде «Рабочее место диалектолога» (РМ), для чего необходимо отделить текст информанта от текста собирателя (квадратными скобками); проставить пробелы перед дефисами — например, перед постчастицами и во всех других случаях, если словоформу необходимо разбирать как 2 слова, а не как одно (*г дому-ту, чему-то, дедушко- домовеюшко*); перевести поданную держателем текстов транскрипцию (иногда изготовленную в самостийных шрифтах или с использованием сразу нескольких шрифтов) в единый шрифт юникода; перевести текст в формат txt в кодировке UTF-8.
2) При открытии уже подготовленного текста в РМ автоматически срабатывают специально созданные **детранскрипторы**: детранскриптор-1 переводит Текст-1 в Текст-2 — первоначальную транскрипцию в «облегченную», унифицированную, детранскриптор-2 переводит Текст-2 в Текст-3 — орфографизированный вариант (= орфографический «подстрочник»), необходимый для работы стандартного

грамматического анализатора. Текст-3 необходимо вручную довести до уровня орфографии, хотя и здесь программа предусматривает некоторую помощь размечающему текст диалектологу: при нажатии кнопки ОРФО (на центральной горизонтальной панели) в Тексте-3 специальной программой осуществляется проверка стандартной орфографии, цветной меткой помечаются слова, с орфографией не совпадающие.

Тексты 1, 2 и 3 выровнены, как в Параллельном корпусе НКРЯ.

3) После обработки Текста-3 осуществляется **грамматическая разметка** текста. Кнопка XML, расположенная на центральной горизонтальной панели РМ на уровне ТЕКСТ, осуществляет прогон грамматического анализатора по всему тексту (работает по Тексту-3), после чего требуется не только **устранить** грамматическую **омонимию** (как в Основном корпусе), но и, ориентируясь на Текст-2, отметить **диалектные грамматические особенности** тех лексем, где эти особенности встретились. Т.е. поверх стандартной грамматической разметки в диалектных текстах предусмотрена возможность отмечать диалектные грамматические особенности лексемы. Для этого в РМ внедрены грамматические таблицы по каждой из 5 изменяемых частей речи (глаголы, существительные, прилагательные, местоимения, числительные).

Если в предыдущем проекте (2008–2009 гг.) нами использовались таблицы с реальными диалектными аффиксами, формами и проч., то к сегодняшнему дню мы отказались от навязчивой конкретики, так как при работе с текстами выяснялось, что списки аффиксов всегда оказывались неполными. В связи с этим большую часть текстов, уже подготовленных для размещения на сайте НКРЯ в рамках предыдущего проекта, пришлось переделывать.

Для каждой изменяемой части речи предусмотрена возможность указывать диалектные особенности на следующих уровнях:

глагол: основа — флексия — суффикс — форма — вид — переходность — возвратность — время. Так, например, диалектные особенности инфинитива отмечаются либо указанием на **диалектный суффикс** (*пекчи*), либо **диал. основу** (*трать, жмать*), **диал. форму** (*идтить*). То же с императивом, где диал. особенности могут проявляться либо на уровне **суффикса** (*посодь*), либо на уровне **формы** (*доедь, ехай*). Иногда лучше отметить сразу несколько возможностей — диал.суффикс и диал. основу (*посодь*).

Особенности спряжения можно указывать следующим образом, например:
общее спряж. (*любют*): *любить* … наст.вр. 3 л. мн. ч. + **диал. флексия**
3-е спряж. (*они гулят*): *гулять*… наст.вр. 3 л. мн. ч. + **диал.флексия**

Таким образом, в ДИАЛЕКТНЫЕ ОСОБЕННОСТИ на уровне «флексии» попадут самые разные вещи: ударные окончания без перехода *е* > *о* (*идёшь*), конечное *-ть* в 3 л. (*растёть*), формы без *-т* (*идё*), общее спряж. (*смотрют*), 3 спряж. (*играт*) и проч. Думается, что специалисту в этом достаточно легко разобраться, особенно при возможности учитывать географические фильтры.

Кнопка ФОРМА зарезервирована для случаев, когда в ЛЯ нет соответствий или трудно/невозможно разделить основу и флексию: типа *jo* или *jy* как формы местоимения *её = она,* 3 л.ж.р. Вин. ед., или *ихной, ихний, ихой* (= *их*), *тэй / тый парень* (= *тот*), *оне, оны* (= *они*), сравн. ст. прилаг. / наречий — *первеющий* и проч.

4) Помимо грамматической разметки, для каждого диалектного текста осуществляется **метаразметка**, которая содержит три уровня:

1 — Адрес-сопровождение
2 — Фонетическая
3 — Диалектная текстовая

Пока что отмечаются лишь немногие фонетические диалектные особенности:

(1) в области ударного вокализма: позиционные чередования гласных после мягких согласных на уровне «старого ятя» и <а>;
(2) в области безударного вокализма: оканье и аканье (включая указания на неполное оканье или диссимилятивное аканье);
(3) в области консонантизма: <г> взрывной или фрикативный и
(4) цоканье.

В банке текстов, переданных для размещения на сайт Национального корпуса, многие тексты оказались поданы вовсе не в транскрипции, а в орфографии, и в «фонетической метаразметке» существует помета «Орфографизированная ли запись?» Помета об «орфографизированной записи» будет постоянно присутствовать и на сайте, чтобы пользователь не пытался делать выводы о фонетических особенностях говора на основе текстов, которые первоначально представлены не в транскрипции.

Если текст, поданный в транскрипции (или в орфографии), будет сопровожден **фонетическим** комментарием, предоставленным держателями текстов, эта информация будет доступна пользователю, т. е. будет активна кнопка «Комментарии к тексту». Это касается и любых других комментариев, связанных с **грамматикой** или **семантикой** предоставленных текстов, и, возможно, даже с какой-то **экстралингвистической** информацией (этнографической, этнокультурной). Можно сопровождать тексты фотоматериалами. Предполагается включать также **аудио- и видеосопровождение**.

Диалектная текстовая метаразметка содержит 3 подуровня: жанр (тип) текста; тематика текста; место и время описываемых событий.

ЖАНР (ТИП) ТЕКСТА делится на 4 категории:

- устные нефольклорные тексты
- письменные нефольклорные тексты
- устные фольклорные тексты
- письменные фольклорные тексты

Пока предпочтение с точки зрения включения в корпус отдается устным нефольклорным текстам, хотя и там могут содержаться элементы фольклора, жанры которых отмечаются не в пределах встречаемости, а в пределах всего

текста в целом (в устном рассказе естественно могут оказаться и колыбельные песни, и частушки, и пословицы, поговорки, загадки, и проч.).

В то же время в банке диалектных текстов уже есть как письменные фольклорные тексты (напр., «песенники» или заговоры, записанные самими носителями), так и письменные нефольклорные (письма, мемуары, дневники и проч.).

Распределение текстов по **тематике** и **семантике** осуществляется, но требует доработки.

Мы пытались ввести заинтересованного читателя (а также зрителя и слушателя) в свою «лабораторию», обратить внимание на некоторые, порой неожиданно возникающие сложности работы с текстами; показать, как сейчас обрабатываются диалектные тексты для НКРЯ и как они будут выглядеть на сайте в ближайшее время.

Свободное предоставление в Интернете текстов русских народных говоров, а также их грамматическая, семантическая и метатекстовая характеристика позволит специалистам-диалектологам, другими лингвистам и нелингвистам, филологам, историкам, культурологам, этнографам — и всем, кто интересуется народным русским словом, обращаться к корпусу в самых разных целях: примеры из текстов и сами тексты могут выступать в качестве справочного материала, материала для научной и педагогической работы, демонстрации этнографических, этнокультурных традиций, особенностей русского менталитета и проч. Некоторые тексты предполагается сопровождать звуко- и видеорядом (в случае, когда тексты явились расшифровками аудио- и видеозаписей). В последующем планируется создание серии интерактивных карт с указанием точки на карте, соответствующей данному пункту с демонстрацией запрашиваемого явления на карте в масштабе области / Европейской части РФ / всей России.

Литература

1. *Качинская И. Б.* (2009), Корпус диалектных текстов в Национальном корпусе русского языка: состояние и перспективы, Лексический атлас русских народных говоров (Материалы и исследования), СПб., с. 57–68. (http://www.philol.msu.ru/~ruslang/pdfs/kachinskaya.i.b/19.pdf от 01.03.2014)
2. *Качинская И. Б.* (2011), Диалектный подкорпус НКРЯ. Новый стандарт подачи. Новое рабочее место, Русская устная речь. Материалы международной научной конференции «Баранниковские чтения. Устная речь: русская диалектная и разговорно-просторечная культура общения». Межвузовское совещание «Проблемы создания и использования диалектных корпусов», СГУ, Саратов, с. 245–255.
3. *Летучий А. Б.* (2005), Корпус диалектных текстов: задачи и проблемы (http://ruscorpora.ru/sbornik2005/13letuchy.pdf), Национальный корпус русского языка: 2003–2005, Индрик, Москва. С. 215–232.
4. *Летучий А. Б.* (2009) Диалектный корпус: состав и особенности разметки (http://ruscorpora.ru/sbornik2008/06.pdf), Национальный корпус русского

языка: 2006–2008. Новые результаты и перспективы, Нестор-История, СПб., С. 114–128.
5. *Русская устная речь* (2011), Материалы международной научной конференции «Баранниковские чтения. Устная речь: русская диалектная и разговорно-просторечная культура общения». Межвузовское совещание «Проблемы создания и использования диалектных корпусов», СГУ, Саратов.
6. *Юрина Е. А.* (2011), Томский диалектный корпус: в начале пути, Вестник Томского гос. ун-та. Филология, № 2, С. 58–63 (http://cyberleninka.ru/article/n/tomskiy-dialektnyy-korpus-v-nachale-puti от 01.03.2014).

References

1. *Jurina E. A.* (2011), The first steps of the Tomsk Dialectal Corpus [Tomskij dialektnyj korpus: v nachale puti], Vestnik Tomskogo gosudarstvennogo universiteta. Filologija, 2, p. 58–63, access at: http://cyberleninka.ru/article/n/tomskiy-dialektnyy-korpus-v-nachale-puti, access date 01.03.2014.
2. *Kachinskaja I. B.* (2009), The corpus of dialectal texts within the Russian National Corpus: current state and plans [Korpus dialektnyh tekstov v Natsional'nom korpuse russkogo jazyka: sostojanie i perspektivy], Studies concerning the Lexical altas of Russian dialects [Leksicheskij atlas russkih narodnyh govorov (Materialy i issledovajia), St. Petersburg, p. 57–68, access at http://www.philol.msu.ru/~ruslang/pdfs/kachinskaya.i.b/19.pdf, access date 01.03.2014
3. *Kachinskaja I. B.* (2011), The dialectal subsorpus of the RNC. The new format and linguist's GUI [Dialektnyj podkorpus NKRJA. Novyj standart podachi. Novoe rabochee mesto], Proceedings of the Conference: Spoken Russian. Dialectal and Colloquial Speech Cultures. Workshop: Building and usage of Dialectal Corpora [Russkaja ustnaja rech'. Materialy mezhdunarodnoj nauchnoj konferencii "Barannikovskie chtenija. Ustnaja rech': russkaja dialektnaja i razgovorno-prostorechnaja kul'tura obschenija". Mezhvuzovskoe soveshchanie "Problemy sozdanija i ispol'zovanija dialektnyh korpusov"], Saratov University, Saratov, p. 245–255.
4. *Letuchij A. B.* (2005), Corpus of dialectal texts: tasks and problems [Korpus dialektnyh tekstov: zadachi i problemy], Russian National Corpus 2003–2005 [Natsional'nyj korpus russkogo jazyka: 2003–2005], Indrik, Moscow, p. 215–232.
5. *Letuchij A. B.* (2009), The dialectal corpus: architecture and tagging properties [Dialektnyj korpus: sostav i osobennosti razmetki], Russian National Corpus 2006–2008: New results and trends [Natsional'nyj korpus russkogo jazyka: 2006–2008: Novye rezul'taty i perspektivy], Nestor-Istorija, Saint-Petersburg, p. 114–128.
6. *Spoken Russian* (2011), Proceedings of the Conference: Spoken Russian. Dialectal and Colloquial Speech Cultures. Workshop: Building and usage of Dialectal Corpora [Russkaja ustnaja rech'. Materialy mezhdunarodnoj nauchnoj konferencii "Barannikovskie chtenija. Ustnaja rech': russkaja dialektnaja i razgovorno-prostorechnaja kul'tura obschenija". Mezhvuzovskoe soveshchanie "Problemy sozdanija i ispol'zovanija dialektnyh korpusov"], Saratov University, Saratov.

ИСПОЛЬЗОВАНИЕ ЛАТЕНТНО-СЕМАНТИЧЕСКОГО АНАЛИЗА В ИССЛЕДОВАНИЯХ И МОДЕЛИРОВАНИИ КОГНИТИВНОГО РАЗВИТИЯ ДЕТЕЙ[1]

Соловьев А. Н. (lechat1@mail.ru)

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

В данной работе представлены результаты исследования когнитивного развития детей дошкольного возраста (4–7 лет), основанного на методе латентно-семантического анализа (ЛСА). Экспериментальная часть состоит из трех тестов. В первых двух тестах продемонстрирована ассоциативно-семантическая связь между моделями ЛСА и ответами испытуемых. В третьем тесте показана возможность использования ЛСА для исследования мнемонических способностей у детей. Проведен сравнительный анализ результатов с моделью ЛСА, полученной на корпусе СМИ.

Ключевые слова: латентно-семантический анализ (ЛСА), ассоциативно-семантические связи, моделирование когнитивного развития детей

USING LATENT SEMANTIC ANALYSIS FOR SIMULATING OF CHILDREN'S COGNITIVE DEVELOPMENT

Solovyev A. (lechat1@mail.ru)

St. Petersburg State University, St. Petersburg, Russia

In the 20th century Noam Chomsky formulated the so-called Plato's problem: why is the amount of our knowledge much greater than we can extract from our everyday experience? For example, the vocabulary of preschool children (aged 6–7) averagely increases by 3–8 words every day, and not every word refers to any reality or action (for example, abstract concepts, words carrying "phatic" or uninformative assignment, etc.).

[1] Работа выполнена при финансовой поддержке Российского гуманитарного научного фонда (РГНФ, проект № 11-06-12042в, 2011–2013 гг.)

How does the child recognize each new meaning of the word and its relation to others, or why are new "meanings" formed? We propose a method to simulate associative-semantic relations between words. On the one hand. it eliminates rigid binding of a lexical unit to any cluster, and on the other it presents a complete system of relationships between words.

The paper presents the results of three experiments with cognitive development of 4–7-year-old children using a Latent Semantic Analysis (LSA) that permits comparisons of semantic similarity between pieces of textual information. We used a technique developed by G. Denhière and B. Lemaire. The principal distinctions of our research are that for the first time, the experiments were performed 1) on the Russian language; and on pre-school children. The children were grouped into two categories: 4–5 and 6–7 years, which corresponds to age variability of cognitive development.

Two experiments describe semantic and associative similarity between LSA models and the children's cognitive development. The third experiment describes using LSA to measure the children's semantic memory. The results are compared to children's model data and adults' model data. The computational models are built from the LSA of a multisource child corpus and of an internet mass media corpus.

Our findings confirm that: 1) LSA can be used to simulate a variety of children's cognitive processes; 2) LSA models represent the development of different age groups children's cognitive processes, in particular associative semantic processes and short-term and long-term memory work; 3) this method may be recommended for the comparative study of children's cognitive development, in particular, the development of associative-logical thinking, verbal discourse, the development of memory.

Keywords: Latent Semantic Analysis (LSA), semantic similarity, semantic association, simulating of children's cognitive development

Введение

Одним из вопросов, которым задаются великие мыслители человечества со времен Платона, является гносеологический вопрос о нашей возможности познаваемости мира. В XX веке Ноам Хомский [Chomsky, 1984] сформулировал так называемую проблему Платона (Plato's problem): почему объем знаний отдельного человека намного больше, чем он может извлечь из своего повседневного опыта? Иначе говоря, как информация, получаемая из последовательности относительно небольшой вариативности событий, может корректно использоваться и адаптироваться к потенциально бесконечному числу ситуаций?

Например, лексикон детей дошкольного возраста (6–7 лет) составляет 3,5–4 тысячи слов, в то время как у детей 10–11 лет этот показатель уже порядка 7–15 тысяч слов [Львов, 2002], т. е. в среднем ежедневно увеличивается на 3–8 слов. При этом не всегда денотат имеет свой строго определенный референт, или, другими словами — не каждое слово имеет соотнесение с реально существующими вещами или выполняемыми действиями (например, абстрактные понятия, слова, несущие «фатическую» или неинформативную нагрузку и пр.).

Возникает вопрос: как ребенок определяет каждое новое значение слова и его соотношение с другими значениями или почему образуются новые денотаты («смыслы») и как они соотносятся между собой?

Современные лингвистические и психолингвистические теории поразному отвечают на этот вопрос: генеративистский подход (см., например, Chomsky, 2002; Fodor, 2009) предполагает существование неких врожденных, а коннекционистский (например, Deacon 2003) — приобретенных языковых структур. Но вопрос остается открытым: что это за структуры, как они работают? Или: каковы перспективы моделирования когнитивных процессов?

Работу «смысловых» механизмов концептуально можно сравнить с процессами категоризации или кластеризации (например, [Соловьев, 2008; Черниговская, 2013 (стр. 28)]). При таком подходе возникает проблема определения изначальных концептов или первичных кластеров, их границ и их числа.

В данном исследовании используется метод, позволяющий моделировать ассоциативно-семантические связи между словами, что с одной стороны позволяет отказаться от жесткой привязки лексической единицы к какому либо из кластеров, а с другой представить целостную систему связей между словами.

1. Латентно-семантический анализ

Одним из методов, позволяющих продемонстрировать работу когнитивных механизмов, является метод латентно-семантического анализа (ЛСА). Еще в 1988 г. Вальтер Кинтч предложил интеграционную модель понимания [Kintsch, 1988], основанную на ассоциативных и семантических связях между лексическими единицами. Именно эта модель и была в дальнейшем реализована в методе латентно-семантического анализа.

Обычно ЛСА используется для выявления латентных (скрытых) ассоциативно-семантических связей между термами (словами, н-граммами) путем сокращения факторного пространства термы-на-документы.

Основная идея латентно-семантического анализа состоит в следующем: если в исходном вероятностном пространстве, состоящим из векторов слов, между двумя любыми словами из двух разных векторов может не наблюдаться никакой зависимости, то после некоторого алгебраического преобразования данного векторного пространства эта зависимость может появиться, причем величина этой зависимости будет определять силу ассоциатино-семантической связи между этими двумя словами.

Существуют три основных разновидности решения задачи методом ЛСА:

- сравнение двух термов между собой;
- сравнение двух документов между собой;
- сравнение терма и документа.

В нашем исследовании мы использовали все три разновидности в зависимости от поставленной задачи. Более подробно о методе ЛСА см. [Landauer, 1998; Соловьев, 2008].

Впервые ЛСА был применен для автоматического индексирования текстов и выявления ассоциативно-семантической структуры текста [Deerwester, 1990]. Затем этот метод был довольно успешно использован для представления баз знаний [Landauer, 1997]. Также метод ЛСА нашел широкое применение в построении когнитивных моделей понимания и формирования знания. В работах [Denhière, 2004; Lemaire, 2003] сделана попытка построить модель долговременной и эпизодической (кратковременной) памяти у детей разного школьного возраста на базе детских текстов. Авторы показали, что семантический анализ текстов методом ЛСА может прояснить как некоторые механизмы работы долговременной и эпизодической памяти, так и связного понимания текста.

Еще одно применение ЛСА нашел в моделях представления и проверки знаний. В вышеупомянутой работе [Landauer, 1997] ЛСА был исследован применительно к известной системе проверки знания английского языка TOEFL на студентах. Также данный метод зарекомендовал себя как эффективное средство проверки и оценочного предсказания для обучающего процесса [Wolfe, 1998]. С помощью ЛСА можно оптимизировать метод обучения, находя оптимальную зону в гауссовом распределении векторного пространства множества знаний. Этим же методом можно давать оценку полученных знаний: студенты, чьи предварительные знания недостаточно хорошо перекрываются семантическим векторным пространством текста, считаются недостаточно хорошо обученными данному предмету, и наоборот.

ЛСА является не единственным методом исследования ассоциативно-семантических связей в тексте. Для выявления лексической синонимии и поиска коллокаций широко используют метод взаимной информации (MI & PMI) и обобщенный ЛСА (GLSA), который представляет собой смесь методов взаимной информации и ЛСА [Matveeva, 2005]. Целесообразность использования того или иного подхода зависит от решаемой задачи. Например, при поиске синонимов методы PMI и GLSA демонстрируют большую точность, чем ЛСА, а в ассоциативных тестах успешнее ЛСА и обобщенный ЛСА [Budiu, 2007]. Следует отметить, что ЛСА является более универсальным методом для моделирования когнитивных процессов, т. к. результаты его работы зависят только от обучающего корпуса и самого процесса обучения (выбора сингулярных значений диагональной матрицы, способа формирования веса термов и пр.). При этом «меру ассоциативности» можно получить для любого слова, содержащегося в обучающем корпусе. MI, PMI и GLSA также зависят от обучающих корпусов, но исследования синонимичности или ассоциативности в рамках этих методов ограничивается предварительно составленными экспериментаторами списками: при их изменении или расширении требуется полный пересчет модели.

2. Эксперимент

В наших экспериментах использовалась методика, разработанная G. Denhière и B. Lemaire [Denhière, 2004, 2007; Lemaire, 2001, 2003]. Основным отличием нашего исследования является: 1) впервые исследование было проведено

на русскоязычном материале; 2) впервые эксперименты проводились на детях дошкольного возраста (4–7 лет). В связи с этим возник ряд трудностей: от полуручной работы по составлению корпусов детских текстов, соответствующим разным возрастам, и разработке необходимого для экспериментов программного обеспечения для ЛСА, до работы с самими детьми — очень сложными испытуемыми.

Дети были разделены на две категории 4–5 и 6–7 лет, что соответствует возрастным особенностям когнитивного развития (см., например, экспериментальные работы Г. Р. Добровой [Доброва, 2007] по исследованию усвоения детьми лексической семантики или исследования Т. В. Черниговской и Т. И. Свистуновой [Черниговская, 2008] организации ментального лексикона и формированию грамматических правил).

2.1. Материал

Для проведения экспериментов были собраны текстовые корпуса, соответствующие детям двух возрастных групп 4–5 и 6–7 лет, которые подверглись ЛСА.

Материал корпуса — детская литература для детей разного возраста, включая рассказы, сказки, детские энциклопедии, разговорный материал.

Общий объем корпуса составил:

- для детей 4–5 летнего возраста — около 2373 тыс. словоформ.
- для детей 6–7 летнего возраста — около 2416 тыс. словоформ.

В сумме около 4789 тыс. словоформ.

После предподготовки корпуса проводился латентно-семантический анализ с разными параметрами. Варьировались количество сингулярных значений диагональной матрицы, расчет весов термов, использование фонетических слов[2], разбиение корпуса на составные части (документы).

В итоге были получены несколько десятков моделей, из которых методом классификации[3] были выбраны две лучшие модели соответствующие двум возрастным группам.

Для сравнения экспериментальных данных с моделью ЛСА «взрослых» текстов был собран корпус СМИ, объем которого составил более 36 млн. словоформ. Источником для корпуса послужили различные официальные интернет-информационные агентства, такие как www.rbc.ru, www.utro.ru, www.rian.ru, www.interfax.ru и др.

Корпус был обработан по аналогичному алгоритму корпуса детских текстов, и были построены несколько моделей ЛСА с разными размерами векторного

---

[2] Фонетическое слово — знаменательная часть слова плюс клитика [Крылов, 2006].

[3] К сожалению, не существует хорошо разработанных методов определения качества получаемой модели: обычно эмпирическим путем отбирают модель, показавшую согласно дизайну эксперимента наилучшие результаты на тестовых данных. Мы использовали метод автоматической классификации текстов, тематически однородных с обучающей выборкой и заранее размеченных экспертом на классы. Для экспериментов отбиралась та модель, которая показывала наилучшие результаты классификации.

пространства и различным количеством факторов, из которых также методом классификации была выбрана наилучшая.

Сокращение сингулярных значений диагональной матрицы при ЛСА составило около 96% при разбиении текстового пространства на документы по 100 предложений. Несколько лучшие результаты показали модели, в которых использовались фонетические слова и веса рассчитывались на основе меры TFiDF.

2.2. Дизайн эксперимента

На основе собранного и обработанного материала был разработан дизайн эксперимента, состоящий из трех тестов:

- «словарный» тест;
- «ассоциативный» тест;
- тест «на память».

2.2.1. Тест первый: «словарный»

Задача: определить связь между семантическим словарем детей разного возраста (две группы) и ЛСА.

Метод: был составлен список из 20 вопросов, к каждому из которых (каждому ключевому слову) были написаны ответы по следующей градации: точный ответ, близкий, слабо связанный или несвязанный ответ. Испытуемые должны были расположить ответы в порядке точности.

Ключевые слова выбирались из корпуса с весами, полученными методом сравнения терма и документа на модели ЛСА.

Было выявлено количество правильных ответов на каждый вопрос, после чего были построены графики зависимости процента правильных ответов от степени точности ответа.

Аналогичные расчеты были проведены на модели ЛСА, полученной из корпуса текстов СМИ.

Пример. Слово: *фокусник.* Ребенку читали слово и просили дать его определение. «Я тебе назову слово, а ты скажи мне, что оно означает».

1. *показывает фокусы* (точный) = .593
2. *показывает сказки* (близкий) = .573
3. *показывает мультфильмы* (слабосвязанный или несвязанный) = .55

2.2.2. Тест второй: «ассоциативный»

Задача: выявить ассоциативные связи между словами у детей двух возрастных групп и сравнить их с результатами ЛСА.

Метод: испытуемым предлагалось 21 слово (стимул), 7 из которых являлись существительными, 7 — глаголами, 7 — прилагательными. Для каждого из них испытуемые называли от 3 до 5 синонимов-ассоциаций. Полученные частотные вектора объединялись и отсортировывались по частоте; после чего вектора сравнивались с ЛСА-вектором.

Стимулы выбирались из материалов полученного корпуса с величинами, полученными из ЛСА методом парного сравнения термов. Рассчитывалось скалярное произведение между каждым стимулом и каждым ответом на стимул, а также среднее как по каждой части речи, так и в целом по группе. Дополнительно к этому рассчитывалось скалярное произведение между стимулом и первыми тремя наиболее частотными ответами на него. Расчет производился на моделях ЛСА, соответствующих каждой возрастной группе.

Аналогичные расчеты проведены на модели ЛСА, полученной из корпуса текстов СМИ.

Пример. Стимул: *море*. Ребенку читали слово и давали задание подобрать к нему несколько ассоциаций. «Я тебе назову слово, а ты постарайся подобрать такие слова, которые с ним могут быть связаны».

Ассоциации приведены в порядке называния со значениями параметров сравнения с вектором ЛСА:

купаться = .827
киты = .729
волнуется = .613
медузы = .457

2.2.3. Тест третий: «на память»

Задача: исследование памяти с помощью моделей ЛСА.

Метод: испытуемому в соответствии с его возрастной группой читался один из двух взятых из корпуса детских текстов типов рассказов. Испытуемый должен был пересказать его в нескольких предложениях. Пересказ записывался на диктофон и был транскрибирован. Пауза между прочтением и пересказом варьировалась от 15 минут до недели. Результаты пересказов сравнивались с ЛСА-моделью методом сравнения документов.

Проведены аналогичные расчеты на модели ЛСА, полученной из корпуса текстов на материале СМИ.

Для проведения тестов был разработано программное обеспечение: 1) программа для построения моделей ЛСА; а также 2) интерфейсная программа для представления результатов моделей латентно-семантического анализа.

2.3. Проведение экспериментов

В исследовании приняли участие 66 детей дошкольного возраста двух возрастных групп: 39 детей 4–5 лет и 27 детей 6–7 лет.

Запись проводилась в два этапа. На первом этапе нашего исследования были протестированы 31 ребенок (16 детей 4–5 лет и 15 детей 6–7 лет). На втором этапе дополнительно к предыдущему было записано еще 35 детей (23 ребенка 4–5 лет и 12 детей 6–7 лет). Основное отличие второго этапа от первого заключалось в том, что пауза между прочтением и пересказом в тесте «на память» составила сутки и более (на первом этапе этот промежуток был 15–30 минут, что не привело к разнице результатах). При этом в тесте

«на память» повторно удалось записать только 31 ребенка (22 — 4–5 лет и 9 детей 5–6 лет).

Тестирование детей проводили в домашних условиях в течение 1–2 часов на каждого ребенка.

В ходе тестирования с использованием «ассоциативного» теста не все дети (в особенности дети первой группы) понимали поставленную перед ними задачу, поэтому детям приводили несколько примеров или задавали наводящие вопросы. Дети называли от одной до четырех ассоциаций; к некоторым словам дети затруднялись подобрать ассоциации.

Тест «на память» вызвал наибольшие трудности, которые связаны с возрастными особенностями детей, а также проблемой обучения пересказу.

3. Результаты

Тест первый: «словарный»

В этом тесте считались средние значения скалярных произведений между словом-стимулом и первыми тремя словами-ассоциациями, последовательно сказанными испытуемыми. Значения усреднялись как по частям речи, так и учитывалось среднее в целом (см. Табл. 1 и Табл. 2).

Табл. 1. Результаты «словарного» теста для детей первой группы (4–5 лет): для каждого слова-ассоциации считалось скалярное произведение с соответствующей моделью ЛСА. В скобках приводится стандартное отклонение.

Стимул	Номер слова-ассоциации		
	1	2	3
Существительные			
голова	0,791615	0,808147	0,791750
дом	0,624179	0,616711	0,597182
лес	0,715897	0,707528	0,726091
подарки	0,530280	0,512679	0,564222
сад	0,710000	0,655471	0,588852
шкаф	0,532842	0,511568	0,468538
яйцо	0,527514	0,553444	0,590296
среднее по классу	0,633190	0,623650	0,618133
Глаголы			
бежать	0,631778	0,667519	0,696579
брать	0,470103	0,549844	0,478826
жить	0,801375	0,702257	0,734148
кричать	0,642892	0,678710	0,704526
научить	0,698686	0,591900	0,623450
плавать	0,533795	0,506622	0,499556
сесть	0,673263	0,630471	0,640913
среднее по классу	0,635984	0,618189	0,625428

Стимул	Номер слова-ассоциации		
	1	2	3
Прилагательные			
быстрый	0,449308	0,501156	0,447407
весёлый	0,621595	0,659667	0,667769
деревянный	0,694333	0,539829	0,618143
красный	0,578450	0,583789	0,571385
маленький	0,596769	0,603636	0,554720
простой	0,663917	0,679929	0,666706
смелый	0,597378	0,594100	0,549650
среднее по классу	0,600250	0,594587	0,582254
среднее по всем	**0,623141 (0,01980)**	**0,612142 (0,01540)**	**0,608605 (0,02310)**

Табл. 2. Результаты «словарного» теста для детей второй группы (6–7 лет): для каждого слова-ассоциации считалось скалярное произведение с соответствующей моделью ЛСА. В скобках приводится стандартное отклонение.

Стимул	Номер слова-ассоциации		
	1	2	3
Существительные			
волшебник	0,748926	0,629727	0,504533
город	0,731120	0,623320	0,613905
за́мок	0,799889	0,732538	0,632875
лето	0,726407	0,678231	0,645652
море	0,646160	0,618826	0,594952
семья	0,752556	0,600115	0,641920
шкаф	0,561593	0,552640	0,498381
среднее по классу	0,709521	0,633628	0,590317
Глаголы			
бежать	0,748905	0,678826	0,683632
болтать	0,732222	0,747269	0,737588
жить	0,785815	0,76284	0,758550
кричать	0,686423	0,697261	0,706385
плавать	0,727958	0,674833	0,566000
сидеть	0,583444	0,566583	0,736550
учить	0,634808	0,585875	0,551174
среднее по классу	0,699939	0,673355	0,677125
Прилагательные			
весёлый	0,765654	0,719708	0,731500
деревянный	0,591462	0,508080	0,610375
красный	0,605462	0,645304	0,589619
маленькая	0,557846	0,565500	0,642625
прекрасная	0,820074	0,772000	0,745381

Стимул	Номер слова-ассоциации		
	1	2	3
сильный	0,686038	0,627826	0,633056
смелый	0,577308	0,536920	0,500579
среднее по классу	0,657692	0,625048	0,636162
среднее по всем	**0,765654 (0,01980)**	**0,719708 (0,01540)**	**0,731500 (0,02310)**

Как показали вычисления, средние (по всем частям речи) значения скалярных произведений у детей второй группы больше по абсолютному значению и имеют более резкий спад от первой к третьей ассоциации, чем у детей первой группы (см. Рис. 1 и Рис. 2). Причем, у обеих групп класс существительных имеет более стабильную тенденцию к спаду.

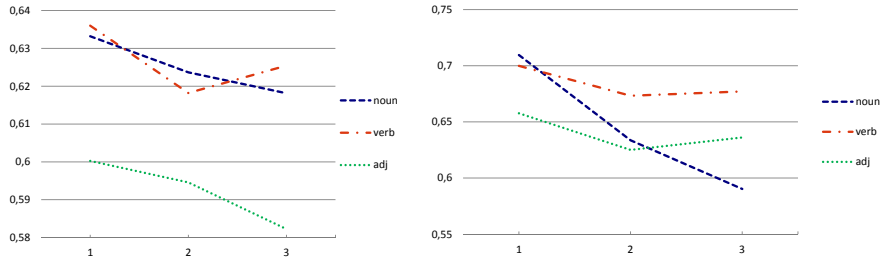


Рис. 1. Средние значения скалярных произведений для первых трех слов-ассоциаций с моделью ЛСА по частям речи для детей первой группы (4–5 лет) — слева и для детей второй группы (6–7 лет) — справа. По оси абсцисс — номера первых трех слов, по оси ординат — значения скалярного произведения.

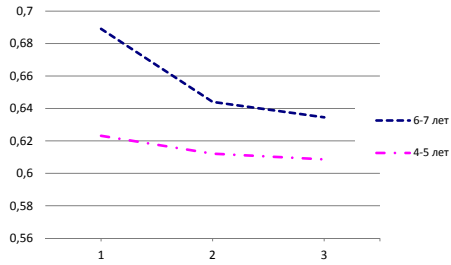


Рис. 2. Средние значения скалярных произведений для первых трех слов-ассоциаций с моделью ЛСА для детей первой группы (4–5 лет) и для детей второй группы (6–7 лет). По оси абсцисс — номера первых трех слов, по оси ординат — значения скалярного произведения.

Аналогичные расчеты были проведены на модели, полученной на текстах СМИ.

Табл. 3. Результаты «словарного» теста для детей первой группы (4–5 лет): для каждого слова-ассоциации считалось скалярное произведение с соответствующей моделью ЛСА, полученной из корпуса СМИ. В скобках приводится стандартное отклонение.

Стимул	Номер слова-ассоциации		
	1	2	3
существительные (среднее по классу)	0,295159	0,266668	0,232557
глаголы (среднее по классу)	0,383547	0,387877	0,372629
прилагательные (среднее по классу)	0,275848	0,220846	0,210666
среднее по всем	**0,318185 (0,051200)**	**0,291797 (0,032400)**	**0,271951 (0,015400)**

Табл. 4. Результаты «словарного» теста для детей второй группы (6–7 лет): для каждого слова-ассоциации считалось скалярное произведение с соответствующим вектором из модели ЛСА, полученной из корпуса СМИ. В скобках приводится стандартное отклонение.

Стимул	Номер слова-ассоциации		
	1	2	3
существительные (среднее по классу)	0,360405	0,333303	0,304718
глаголы (среднее по классу)	0,373503	0,369079	0,346461
прилагательные (среднее по классу)	0,300056	0,294374	0,291618
среднее по всем	**0,344654 (0,031500)**	**0,332252 (0,027500)**	**0,314266 (0,009200)**

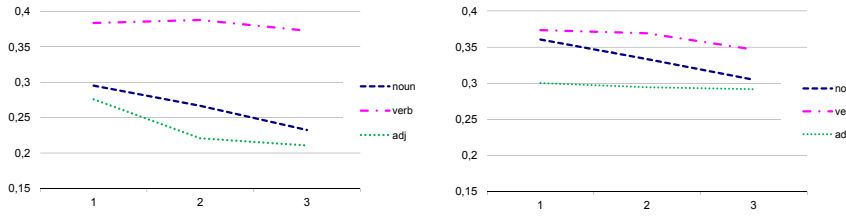


Рис. 3. Средние значения скалярных произведений для первых трех слов-ассоциаций с моделью ЛСА, полученной из корпуса СМИ, для детей первой группы (4–5 лет) — слева и для детей второй группы (6–7 лет) — справа. По оси абсцисс — номера первых трех слов, по оси ординат — значения скалярного произведения.

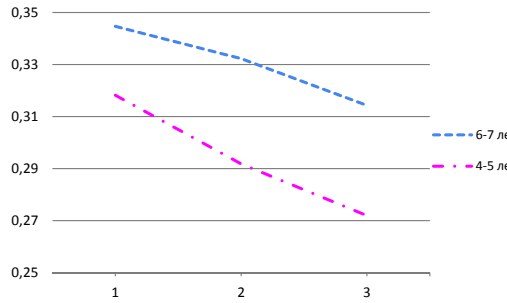


Рис. 4. Средние значения скалярных произведений для первых трех слов-ассоциаций с моделью ЛСА, полученной из корпуса СМИ, для детей первой группы (4–5 лет) — слева и для детей второй группы (6–7 лет) — справа. По оси абсцисс — номера первых трех слов, по оси ординат — значения скалярного произведения.

Как видно из рисунков 3 и 4, значения скалярных произведений у детей первой и второй групп также имеют в среднем тенденцию к уменьшению с каждым последующим ответом, при этом абсолютные значения скалярных произведений примерно в два раза меньше.

Частотный анализ слов-ассоциаций и значения скалярного произведения с моделью ЛСА нескольких первых наиболее частотных слов не показал стабильной зависимости.

Тест второй: «ассоциативный»

Для каждой группы (4–5 лет — 39 испытуемых и 6–7 лет 27 испытуемых) детей было подсчитано количество данных ответов на каждую группу ассоциаций: выбор одного из трех вариантов (точного, близкого и слабо связанного или несвязанного) означал единицу, остальное ноль (Табл. 5).

Табл. 5. Сумма выбранных вариантов ответов детей двух групп (в абсолютном и процентном соотношении); 1 — точный, 2 — близкий, 3 — слабо связанный или несвязанный ответ.

	4–5 лет		6–7 лет	
1	490	63,06%	424	78,37%
2	148	19,05%	99	18,30%
3	139	17,89%	18	3,38%
сумма	777	100%	541	100%

Анализ выбора вариантов показал, что дети второй группы значительно лучше выбирают наиболее близкий ассоциативный вариант (Рис. 5, 6).

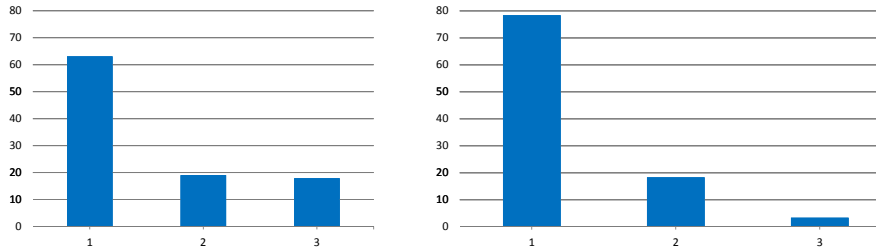


Рис. 5. Результирующие графики выбранных ответов детей первой (слева) и второй (справа) групп. По оси абсцисс — варианты ответа, по оси ординат — сумма ответов в процентном соотношении.

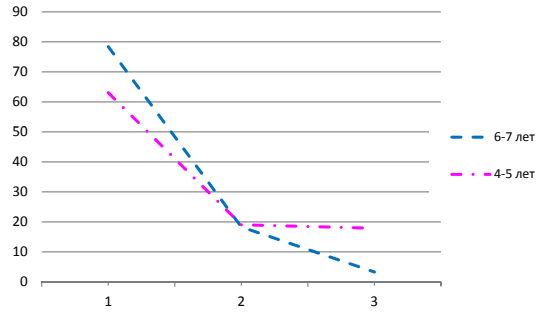


Рис. 6. Линии тренда ответов для детей первой и второй групп. По оси абсцисс — варианты ответа, по оси ординат — сумма ответов в процентном соотношении.

Для обеих групп испытуемых был посчитан коэффициент согласованности (Coeff. of Concordance) ответов. Для первой группы он составил 0,24853, для второй группы — 0,57919.

Далее были получены средние значения результатов значений косинуса между вопросом и ответом для каждой возрастной группы. Аналогичная процедура была проведена с результатами значений косинуса, полученных из модели СМИ.

Результаты в абсолютном и процентном отношении представлены в таблице № 6.

Табл. 6. Средние значения косинусов на «детских» и «взрослом» корпусах в абсолютном и процентном соотношении (в скобках)

Ответы	**1**	**2**	**3**
4–5 лет			
модель 4–5 лет	0,7708 (38,45%)	0,68705 (34,27%)	0,5467 (27,27%)
модель СМИ	0,5126 (34,0%)	0,4958 (32,89%)	0,49885 (33,10%)
6–7 лет			
модель 6–7 лет	0,81175 (37,75%)	0,73395 (34,13%)	0,6045 (28,12%)
модель СМИ	0,4813 (33,57%)	0,4826 (33,66%)	0,4699 (32,77%)

Как видно из таблицы, в отличие от «детских» моделей на «взрослой» модели значение угла косинуса почти не меняется.

Тест третий: «на память»

В третьем тесте измерялась величина скалярного произведения текстов: реального и пересказанного испытуемым в разные моменты времени. На втором этапе в этом тесте был увеличен промежуток времени после первого пересказа до суток и более. Это было связано с тем, что промежуток записи в 15 минут не привел к значимой разнице при анализе данных.

В силу ряда причин, не зависящих от исследователей, повторно удалось записать не всех детей. В итоге в данном тесте были заново записаны 31 ребенок 4–5 лет и 10 детей 6–7 лет.

Результаты сравнения по мере косинуса двух текстов пересказа с источником представлены в табл. 7.

Табл. 7. Значения скалярного произведения исходного и пересказанных детьми текстов за разный промежуток времени

	повтор сразу	**через сутки и более**
4–5 лет		
модель 4–5 лет	0,482375	0,435139
модель СМИ	0,485271	0,467736
6–7 лет		
модель 6–7 лет	0,489400	0,488600
модель СМИ	0,512300	0,514500

Как видно из таблицы 7, значения меняются только у испытуемых первой группы как на «детской», так и на «взрослой» модели.

Табл. 8. Коэффициенты корреляции между результатами ответов сразу и через промежуток времени для испытуемых обеих групп

	коэф.корреляции
модель 4–5 лет	0,326179
модель СМИ	0,461313
модель 6–7 лет	0,350000
модель СМИ	−0,418414

В таблице 8 представлены результаты расчета коэффициента корреляции ($p < 0,05$) между результатами ответов двух этапов теста «на память» (сразу и с задержкой) для всех моделей ЛСА. Как видно из таблицы корреляция результатов лучше для моделей ЛСА, полученных на корпусе СМИ.

4. Выводы

Анализ результатов «ассоциативного» теста показал, что средние значения скалярных произведений первых трех ассоциаций у детей второй, более старшей группы больше по абсолютному значению и имеют тенденцию к более резкому снижению при ослаблении ассоциации (Рис. 2). По всей видимости, это связано с тем, что у детей 6–7 лет должна быть больше развита ассоциативность по сравнению с детьми 4–5 лет. Таким образом, можно сделать вывод, что соответствующие возрастам модели адекватно отражают ассоциативно-семантические связи для когнитивного развития детей обеих групп.

Распределение слов-ассоциаций по частям речи не выявило статистических особенностей в виду небольшой статистики (максимальная частота слова была около 6–7 для стимула, а зачастую и меньше, что недостаточно для статистического анализа, стабильность которого определяется десятками и сотнями повторений).

Сравнение с моделью СМИ показало похожие результаты. При этом абсолютные значения косинуса получились почти в два раза меньше, чем результаты, полученные на «детских» моделях. Это говорит о том, что ассоциативность «взрослой» и «детских» моделей имеют одинаковую тенденцию к ослаблению. При этом «взрослая» ассоциативность на «детских» корпусах ожидаемо ниже, т. к. смоделирована на других текстах. Другой особенностью данного сравнения является то, что если «детские» модели показали разную линейность (разную скорость ослабления ассоциативности) в зависимости от возрастной группы, то на «взрослом» корпусе кривые практически параллельны. Это подтверждает то, что соответствующие «детские» модели более корректно, чем «взрослая» модель, описывают ассоциативно-семантические связи для детей двух возрастных групп.

Результаты «словарного» теста уверено подтверждают результаты «ассоциативного» теста: модели ЛСА согласуются с данными когнитивного развития детей. У испытуемых старшего возраста более развиты ассоциативно-синонимические понятия. Это видно по линии тренда: у детей 6–7 лет более сильная зависимость от силы ассоциативности стимулов (более крутой спад линии тренда), в то время как у детей 4–5 лет связь менее сильна и нелинейна (Рис. 6). Этот результат подтверждает коэффициент согласованности: у детей 6–7 лет он примерно в два раза выше.

Сравнение полученных результатов с «взрослой» моделью ЛСА показало, что значения косинуса на модели СМИ практически не меняется от связанности ответа (Таб. 6). Причиной этого может быть то, что во «взрослом» корпусе отсутствует часть лексики «детских» корпусов, поэтому сравнение идет, в основном, по наиболее употребимым словам, которые примерно одинаково распределены в ответах.

В любом случае сравнение результатов ответов детей разного возраста как между собой, так и со «взрослой» моделью говорит о том, что 1) «детские» модели различаются по силе ассоциативности: у детей старшего возраста ассоциативность развита лучше; 2) модели взрослого и детского восприятия на лексическом уровне существенно различаются.

Исследование памяти (тест «на память») на данных моделях продемонстрировало результаты только для детей первой возрастной группы. Причем

это видно как на «детской», так и на «взрослой» модели, хотя и не столь выраженно. Результаты второй группы не показали изменений. Возможно, это связано с тем, что, во-первых, более старшие дети уже лучше запоминают тексты, а во-вторых, возможно это связано с неудачно выбранной методикой экспериментов: в виду объективных причин дети опрашивались в разные промежутки времени (на следующий день, через 5–7 дней), что не учитывалось при обработке; к тому же детей этой группы удалось записать в три раза меньше, чем первой. Исходя из этого, можно сказать, что метод ЛСА может быть использован при исследовании как кратковременной, так и долговременной памяти у детей, но это требует проверки дальнейшими экспериментами.

Таким образом, результаты нашего исследования показывают, что:

- Метод ЛСА может быть использован для исследований когнитивного развития детей.
- Используемые модели ЛСА отображают процессы развития когнитивного развития детей разных возрастных групп, в частности ассоциативно-семантические процессы и работу кратковременной и долговременной памяти.
- Данный метод рекомендуется использовать для сравнительного исследования когнитивного развития детей, в частности, развития ассоциативно-логического мышления, речевого дискурса, развития памяти.

Литература

1. *Доброва Г. Р.* О некоторых аспектах усвоения лексической семантики детьми 3–6 лет: влияние «нового знания» на речевое поведение // Возраст как фактор речевого поведения. Сборник статей. Пермь: Изд-во ГОВПО Пермского гос. ун-та, 2007.
2. *Крылов С. А.* Делимитация тактов в русском письменном тексте // Труды международной конференции «Корпусная лингвистика-2006». СПб.: Изд-во СПбГУ, 2006. — С. 54–55.
3. *Львов М. Р.* Основы теории речи. — М., 2002.
4. *Соловьев А. Н.* Язык, мышление и современные системы понимания речи // Вестник СПбГУ. Серия Биология (3). Вып. 1. СПб., Изд-во СПбГУ, 2008. С. 99–104.
5. *Соловьев А. Н.* Моделирование процессов понимания речи с использованием латентно-семантического анализа / Диссертация на соискание степени к. ф.н. СПбГУ, 2008.
6. *Черниговская Т. В.* Чеширская улыбка кота Шрёдингера: язык и сознание. — М.: Изд. Языки славянской культуры, 2013.
7. *Черниговская Т. В., Гор К., Свистунова Т. И.* Формирование глагольной парадигмы в русском языке: правила, вероятности, аналогии как основа организации ментального лексикона (экспериментальное исследование) // Когнитивные исследования. Сб. научн. трудов. Вып. 2. / Отв. ред. Т. В. Черниговская, В. Д. Соловьев. М.: Изд-во «Институт психологии РАН», 2008. С. 165–181.

8. *Budiu, R., Royer, C., & Pirolli, P. L.* Modeling information scent: a comparison of LSA, PMI and GLSA similarity measures on common tests and corpora. Proceedings of the 8th Annual Conference of the Recherche d'Information Assistée par Ordinateu. Paris, France, 2007– P. 314–332.
9. *Chomsky N.* Modular Approaches to the Study of the Mind. San Diego: San Diego State University Press, 1984.
10. *Chomsky N.* New Horizons in the Study of Language and Mind. Cambridge University Press, 2002.
11. *Deacon T. W.* Multilevel selection in a complex adaptive system: The problem of language origins. Weber B., Depew D. (eds.). Evolution and Learning: The Baldwin Effect Reconsidered. MIT Press. 2003.
12. *Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R.* Indexing by Latent Semantic Analysis // Journal of the American Society for Information Science. 1990. 41(6). — P. 391–407
13. *Denhière G., Lemaire B., Bellissens C., Jhean-Larose S.* A semantic space for modeling children's semantic memory // D. McNamara, T. Landauer, S. Dennis, W. Kintsch (eds.). The handbook of Latent Semantic Analysis. Mahwah: Lawrence Erlbaum Associates, 2007. — P. 143–165.
14. *Denhière G., Lemaire B., Bellissens C., Jhean-Larose S.* Psychologie cognitive et compréhension de texte: une démarche théorique et expérimentale // S. Porhiel, D. Klinger (eds.). L'unité texte. Pleyben: Perspectives, 2004. — P. 74–95.
15. *Fodor J.* Where is my mind? London Review of Books. Vol. 31, № 3, 12 February. — 2009.
16. *Landauer T. K., Dumais S. T.* A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge // Psychological Review. 1997. 104. — P. 211–240.
17. *Landauer T. K., Foltz P., Laham D.* An Introduction to Latent Semantic Analysis. Discours Processes, 25, 1998 — P. 259–284.
18. *Lemaire B., Bianco M., Sylvestre E., Noveck I.* Un modèle de compréhension de textes fondé sur l'analyse de la sémantique latente // La cognition entre individu et société (actes du colloque de l'ARCo) / H. Paugam Moisy, V. Nyckees, J. Caron-Pargue (eds.). Hermès, 2001. — P. 309–320.
19. *Lemaire B., Denhière G.* Cognitive Models based on Latent Semantic Analysis // Tutorial given at the 5th International Conference on Cognitive Modeling (ICCM'2003), Bamberg, Germany, April 9 2003. — P. 23–25.
20. *Matveeva, I., Levow, G., Farahat, A., & Royer, C.* Terms representation with generalized latent semantic analysis. In Proc. ranlp 2005.

ASSOCIATING SYMPTOMS WITH SYNDROMES. RELIABLE GENRE ANNOTATION FOR A LARGE RUSSIAN WEBCORPUS

Sorokin A. (alexey.sorokin@list.ru),
Katinskaya A. (a.katinsky@gmail.com),
Sharoff S. (s.sharoff@leeds.ac.uk)

Lomonosov Moscow State University,
Russian State University for the Humanities, Moscow, Russia;
University of Leeds, Leeds, UK

The paper describes several experiments aimed at establishing the parameters for genre annotation of potentially any text which can be collected from the Russian web. We started with a set of text classification parameters, refined them iteratively in several studies and established a reliable framework, which was further subjected to clustering analysis. Overall, we obtained the level of agreement for Krippendorff's α to be in the range of $0.51<\alpha<0.84$. We have also discovered the most common combinations of parameters in the test corpus, which should form the basis for classifying very large samples of the Russian web.

Keywords: genre annotation, webcorpora, reliability of annotation, clustering

1. Introduction

Genre classification is often referred to as "jungle", this metaphor has been used by numerous researchers (Kilgarriff, 2001; Lee, 2001; Sharoff, 2010). While the web users see the differences between individual texts, it is difficult to agree upon a set of labels, which can cover the majority of webpages and which can be, at the same time, applied reliably by annotators (Sharoff et al., 2010). One reason for this is the sheer number of possible genre labels, up to 6,000 according to some studies (Adamzik, 1995). Another reason is a high degree of genre hybridism, especially on the web where many texts are not controlled by the institutional gate-keepers (Santini et al., 2010).

This study continued a line of genre classification experiments which started from adaptation of John Sinclair's typology of communicative aims to the needs of the Russian National Corpus (Sinclair, 2003; Sharoff, 2005), which in turn led to the Functional Genre Classes used in (Sharoff, 2010). The lack of reliability in classification of random web texts, which we investigated in the TTC project[1], led to a proposal

[1] http://www.ttc-project.eu/

for introducing Functional Text Dimensions (FTDs), which can be used to judge the similarity of texts (Forsyth and Sharoff, 2014). It proposed such FTDs as:[2]

A1: **Argumentative** To what extent does the text argue to persuade the reader to support (or renounce) an opinion or a point of view?

A7: **Instructive** To what extent does the text aim at teaching the reader how to do something?

A12: **Promotional** To what extent does the text promote a commercial product or service?

The annotators can express their opinions concerning each of these parameters on a scale from 0 (absent) to 2 (strongly present). From earlier studies we know that the annotators tend to achieve better agreement using such scales in comparison to atomic genre labels (Forsyth and Sharoff, 2014; Sharoff et al., 2010). The higher values in FTDs then serve as "symptoms", from which we can infer a number of "syndromes", i.e. traditional look'n'feel genres.

In this study we try to investigate the FTD values obtained from a diverse range of Russian texts to improve the FTD set and also to compare it against a recent project aimed at wide-ranging genre classification of the web (Egbert and Biber, 2013). The goal of this study is to make genre annotation reliable, i.e. the annotators should not struggle with the ambiguity of their choices, interpretable, i.e. the investigators understand the ranges of their choices, and applicable to any web page containing running text.

In Section 2 we introduce the corpus and the annotation procedure. In Section 3 we present the level of the interannotator agreement achieved on our corpus. In Section 4 we describe experiments at clustering our annotated corpus using the FTD values as an attempt to map the "symptoms" observed in our annotation to "syndromes". Finally in Section 5 we analyse the results by comparing the clusters against the register list from (Egbert and Biber, 2013), which is in turn based on an earlier study (Rehm et al., 2008).

2. Corpus and the Annotation Procedure

Corpus selection and annotation went through three stages.

Stage 1

We selected 226 texts from Open Corpora (Bocharov et al., 2011) with the intention of enriching this collection with genre labels. Using texts with permissible licenses should have also helped in distributing the results of annotation freely. We requested two annotations per text to determine the degree of inter-annotator agreement. Another goal was to "debug" the definitions of the Functional Text Dimensions. For independent attribution of texts we used the Brown Corpus categories.

[2] See the full FTD list which we use in Appendix 1.

Stage 2

Since we found that the genre range of the Open Corpora collection is quite limited, we enriched the corpus by requesting *diverse* genre sets from a new round of annotators, who were also asked to annotate their own texts. The main sources of the texts were blogs, news portals, Wikipedia and other online encyclopedias, forums, online magazines, web libraries, promotional web-sites and legal resources. This iteration led to further feedback on the FTDs. We decided to add three more FTDs such as A16 (defining a topic), A17 (expressing judgments) and A18 (interaction between participants). This was primarily due to the fact that some texts in Corpus 2, e.g. Wikipedia articles, were not distinguished from other texts by the fifteen original FTDs.

Stage 3

Corpus 2 consisting of 514 texts was annotated by 11 annotators, overall 3 annotators per text. Table 1 describes quantitative characteristics of Corpus 1 and Corpus 2. Given that the Brown Corpus categories were found to be unsuitable for a large number of webtexts, we asked the annotators to attribute each text by using one of the registers used in Egbert and Biber (2013), e.g. 'News report/blog', 'Description with intent to sell', 'Review'. This was needed to get some attribution which is independent from the FTDs and to evaluate the results. In addition, the annotators were asked to give comments about the annotation procedure, the FTDs and the registers which caused the greatest difficulty for annotation.

As we are primarily interested in investigating *language* of the Web, we deliberately focused on the webpages with running text, leaving out such types of webcontent as social network profiles, e-commerce pages, forms, etc.

Table 1. Annotated corpora used in this study

Corpus 1			**Corpus 2**		
Documents	Words	Sentences	Documents	Words	Sentences
226	273,319	16,587	514	640,476	41,961

The annotation results show considerable correlation between several functional dimensions, such as A10 and A18, A14 and A15. Strong correlation between A10 and A18 gives reasons for removing duplicate FTDs. In the case of A14 and A15, we can assume that correlation was influenced by a small number of A15 documents that do *not* belong to the field of Science and Technology. The last annotation round also suggested a new "poetic" FTD, which roughly corresponds to the poetic function of language according to Jakobson (1960), covering various kinds of texts concerned with making an aesthetic impression.

3. Interannotator Agreement

We measured the interannotator agreement using several methods, primarily using Krippendorff's α (Krippendorff, 2004), which treats the annotators are interchangeable

and measures the difference between disagreement expected by chance vs observed disagreement:

$$\alpha = 1 - \frac{D_{observed}}{D_{chance}}$$

The corresponding results are shown in Table 2. Overall, the agreement is above 70%, but for some FTDs, such as newly added *A*16,*A*17 and *A*18 the level of agreement was lower. Some FTDs tested in previous studies, in particular *A*5 (flippant) or *A*6 (informal) also demonstrated considerable disagreement between annotators. However, in comparison to achieving agreement on atomic genre labels (Sharoff et al., 2010), FTDs overall offer better acceptable agreement judgments.[3] We also measured Krippendorff's alpha for the Biber register labels, the agreement was 53.0% for the register labels, 65.8% for the general registers and 58.7% using a compound distance function, assigning 1 for different labels in the same general register and 2 for a pair of elements in diverse registers.

Table 2. Krippendorff's α values for FTDs and annotators

	A1	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
Overall	56.39	58.44	71.80	51.39	55.53	69.58	78.00	84.22
Annotator 1	67.31	51.16	83.86	58.42	35.33	69.15	84.62	79.66
Annotator 2	54.67	68.94	71.71	35.06	61.41	68.86	64.68	80.56
Annotator 3	28.81	53.22	68.60	58.50	52.36	64.87	77.14	81.71
Annotator 4	66.34	54.51	72.39	60.72	53.66	80.03	58.48	87.19
Annotator 5	58.25	67.92	82.74	36.04	69.28	84.89	77.65	81.16
Annotator 6	50.62	57.37	60.14	40.69	60.92	69.69	61.25	69.63
Annotator 7	66.13	72.92	84.04	41.73	59.13	83.87	82.17	81.01
Annotator 8	50.35	64.07	77.38	57.04	56.40	52.30	85.09	91.08
Annotator 9	58.82	50.67	69.86	62.38	33.24	46.23	38.87	92.96
Annotator 10	38.26	34.99	61.49	46.11	52.11	80.19	88.85	81.30
Annotator 11	51.64	58.51	52.51	45.02	64.61	74.49	74.83	90.38
	A12	**A13**	**A14**	**A15**	**A16**	**A17**	**A18**	**Total**
Overall	62.09	58.71	52.14	56.80	63.55	51.43	49.32	62.95
Annotator 1	62.68	51.42	59.24	61.53	42.80	64.51	34.72	62.99
Annotator 2	46.58	57.11	50.02	62.89	69.89	60.04	34.66	62.78
Annotator 3	86.29	60.07	12.59	57.23	59.54	−12.33	66.03	54.99
Annotator 4	75.61	60.92	66.03	67.73	53.15	50.44	59.79	65.72
Annotator 5	42.25	49.72	60.59	64.84	70.52	48.29	58.08	66.19
Annotator 6	52.29	65.94	53.21	46.59	54.84	46.84	40.27	59.27
Annotator 7	63.86	47.40	57.57	64.07	56.01	62.10	36.93	67.14
Annotator 8	59.46	52.94	46.10	54.98	74.17	57.94	53.34	67.26
Annotator 9	64.96	72.58	10.62	32.36	67.45	58.04	55.90	57.85
Annotator 10	63.00	62.21	61.90	59.55	65.49	33.79	50.67	59.93
Annotator 11	61.35	66.78	40.09	38.53	60.98	48.12	47.08	62.08

[3] $\alpha \geq 60\%$ is usually treated as the acceptability threshold (Krippendorff, 2004).

Given systematic errors from some annotators for some FTDs (caused by their misunderstanding of the instructions), as well as possible issues with fatigue (coming from annotation of more than 100 texts), we need to exclude badly annotated texts from future research since the characteristics of such documents cannot be considered as "predictive" values of the FTDs. The second methodological reason for reducing the number of texts is that it is worse in general to have noisy data than to have less data, especially when the nature of noise is unclear. In the clusterisation task the presence of noise may potentially deform the shape of clusters or create spurious clusters which do not correspond to any pattern in the data. However, we cannot simply exclude from the sample the documents for which there was no agreement since two or more annotators agreed on all the FTDs only for 199 of 500 initial texts. Moreover, we found that the more informal the text is the lower the probability of their agreement. It implies that the distribution of the FTD values over the reliable texts considerably differs from the same distribution over the set of all documents.

Therefore, we tried to preserve as many documents as possible unless they can be considered as reliable. The key idea is to remove the *ratings* of annotators when they disagree with other annotators and keep their ratings otherwise. For every FTD we evaluated the quality of experts using Krippendorff's α: for any text annotated by k annotators (k=3 in our case) we created $k-1$ pairs of ratings assigned by the current annotator together with the rating given by another annotator. Then for every FTD we selected the worst annotator and removed his/her ratings for all documents where s/he disagreed with other annotators. Then individual annotation qualities were recalculated and a new worst annotator was selected until the agreement quality for the worst annotator reached a fixed threshold or became close to the mean value of individual agreement for all annotators. We used the α value of 0.75 for the threshold, also we tried the values from 0.6 to 0.7. Since these thresholds produced a lower number of reliable texts, we allowed different annotators to disagree by 0.5. The sizes of obtained collections of documents are shown in table 3. Table 4 contains the values of Krippendorff's α for FTDs after the removal procedure with threshold 0.75 and 0.5 disagreement allowed.

Table 3. The number of remaining documents for different selection strategies

Selection strategy	# documents
Alpha threshold 0.75	283
Alpha threshold 0.75, 0.5 disagreement allowed	315
Alpha threshold 0.6	238
Alpha threshold 0.6, 0.5 disagreement allowed	263
2 annotators agreement	199
No selection	500

Table 4. Krippendorff's α values for FTDs after removing unreliable annotations

	A1	A3	A4	A5	A6	A7	A8	A9
Overall	86.73	81.34	87.64	76.21	89.88	87.97	90.06	86.28
Best annotator	97.09	92.81	94.46	96.98	96.77	93.93	99.67	93.96
Worst annotator	75.07	73.32	80.06	65.51	78.50	77.75	78.84	80.20

	A12	A13	A14	A15	A16	A17	A18	Total
Overall	88.47	88.70	53.24	58.50	88.47	58.62	85.26	80.97
Best annotator	98.45	93.09	66.89	68.94	96.21	71.56	94.68	81.79
Worst annotator	80.47	75.17	12.57	39.74	75.77	21.81	77.48	77.53

4. Clustering

The ultimate goal of our project is to provide automatic genre classification using the FTDs, which can be detected reliably, and also to map them to genre syndromes. Therefore, we are interested in exploring frequent combinations of FTD values corresponding to stable patterns, which may be considered as similar to genres in a traditional genre system like Egbert and Biber (2013). Our primary task in this section is to detect clusters in the FTDs space and to investigate whether these clusters are linguistically relevant.

We used several methods to cluster the texts by the values of their FTDs. Since the FTD values are discrete, we used the taxicab (also known as Manhattan) metric as the distance function. The discreteness of our feature space makes the clustering task problematic. The general difficulty is that we do not know the probability distribution on the sample set, especially on its subset selected for clusterisation. Also it is difficult to perform feature weighting for discrete features with a small number of feature values. Both agglomerative and iterative methods have their own drawbacks complicating their usage for our task.

The main difficulty with the agglomerative methods is their sensitivity to the order of objects. This sensitivity becomes more dramatic since the discreteness of our feature space creates many ties. Also it is impossible to correct the wrong choices made during early clusterisation steps. However, the usage of iterative methods is even more problematic. First of all, the most common methods such as *EM* or *K*-means are not suitable for discrete spaces. We can use *k*-medoids instead but this does not fix the problem of detecting the number of clusters. Also iterative methods are sensitive to initial cluster approximations.

Therefore we decided to combine hierarchical and iterative methods. First of all we performed agglomerative clustering using the weighted linkage. We detected the number of clusters k searching for the knee position of the evaluation graph by the well-known L-method (Salvador and Chan, 2004). Then we used these clusters as initial approximations for the *k*-medoids algorithm. We also tried *k*-means, even though it has limitations when applied to discrete data.

The proposed method of clusterisation is unstable, since it is sensitive to the initial order of objects. We measured the robustness of clusterisation using the adjusted Rand index (ARI), which measures the degree of intersection between different clusterisation runs. Precisely, we ran the clusterisation algorithm for 20 different random orderings of data and calculated the average value of ARI between the clusters obtained in this way. To estimate the internal quality of clustering we applied the silhouette score (Rousseeuw, 1987), which assesses how well all objects lie within their clusters.

The values of the indices for different selection strategies and different clusterisation methods are shown in Table 5, each vertical section corresponds to a particular data selection method and consists of two rows: the first row contains the values of the robustness index (R), the second one contains the silhouette score (S).

Unfortunately, the scores are not particulary high because of the annotation noise. It means the additional procedure of noise removal must be performed. We determine and remove the outliers using the silhouette score. To be removed an object should have the silhouette score lower then the minimum of the predefined threshold t (we choose t=0.25) and current threshold $t'=\mu-\sigma$ with μ and σ being, respectively, the mean and the deviation of the individual silhouette scores. After excluding the unreliable objects, the scores were recalculated, and the process was repeated until the Rand score exceeds 90%. The convergence of the algorithm was rather fast which means that the core elements of clusters are clustered independently of data ordering.

Table 5. Robustness and silhouette for different selection strategies

Selection method	# documents	*k*-means	*k*-medoids	Hierarchical
0.75	283	0.864	0.766	0.769 (R)
		0.468	0.425	0.423 (S)
0.75, 0.5 disagreement	315	0.810	0.757	0.714 (R)
		0.444	0.431	0.380 (S)
0.6	238	0.844	0.805	0.871 (R)
		0.480	0.440	0.383 (S)
0.6, 0.5 disagreement	263	0.783	0.745	0.719 (R)
		0.427	0.440	0.365 (S)
2 annotators agreement	199	0.890	0.863	0.827 (R)
		0.508	0.505	0.450 (S)
No selection	500	0.691	0.531	0.490 (R)
		0.329	0.318	0.168 (S)

The values of the scores are presented in Table 6. Each vertical section contains four rows, containing the values of the adjusted Rand index, the silhouette score, the number of objects attached to clusters and the final number of clusters.

Table 6. Different methods for cleaning clusters

Selection method	Initial # documents	*k*-means	*k*-medoids	Hierarchical
0.75	283	0.924	0.984	0.972
		0.670	0.660	0.625
		206	194	203
		16	14	15
0.75, 0.5 disagreement	315	0.936	0.993	0.927
		0.566	0.686	0.530
		248	209	219
		8	16	8

Selection method	Initial # documents	*k*-means	*k*-medoids	Hierarchical
0.6	238	0.954	0.925	0.931
		0.631	0.605	0.486
		188	196	206
		9	10	8
0.6, 0.5 disagreement	263	0.959	0.956	0.948
		0.574	0.550	0.579
		215	216	214
		11	11	11
2 annotators agreement	199	0.991	0.991	0.968
		0.696	0.654	0.659
		153	162	160
		13	9	12
No selection	500	0.915	0.953	0.998
		0.414	0.570	0.641
		358	295	250
		13	12	15

So different clusterisation algorithms produce different number of clusters in the final clusterisation experiment. Though these values are very important for further usage of obtained clusters as classes for document classification, they cannot be considered as absolutely reliable. To uncover better the cluster structure we made an additional experiment when the number of clusters was fixed through the algorithm. We varied the number of clusters from 8 to 15. In this experiment we used the initial set of 315 documents, obtained in the run of selection algorithm with 0.75 threshold and allowed 0.5 disagreement between annotators.

Some basic clusters were found independently of the number of clusters. The documents in each cluster usually share principal FTDs, i.e. FTDs, which are equal to 2. Below we list these clusters together with their principal dimensions.

1pt 0pt

1. "Instructions" (21 texts), principal dimension *A*7.
2. News (64 texts), principal dimension *A*8.
3. "Legalese"(11 texts), principal dimension *A*9.
4. "Specialised technical texts" (13 texts), principal dimensions *A*14,*A*15.
5. Descriptive and encyclopedic texts (49 texts), principal dimension *A*16.
6. Adverts (13 texts), principal dimensions *A*1,*A*12,*A*17.
7. Argumentative propaganda texts (13 texts), principal dimensions *A*1,*A*13,*A*17.

5. Interpretation of Clustering Results

We were interested in comparing the clusters to Biber's registers. We identify the clusters with the set of their principal FTDs, while indicating other FTDs which are significant only for a portion of our texts.

The cluster with the principal dimension A9 is the only cluster which contains documents annotated within a single Biber's register ('Legal terms and conditions'). Some texts of the A9 cluster also contain higher values of the A7 FTD (instructions), describing a proper legal procedure to achieve something, as well as, the A16 FTD (a text defining a topic), e.g., definitions of the kinds of taxation or of land property.

The cluster with the principal dimension A7 includes several kinds of instructional texts according to Biber ('How-to', 'Instructions', 'Recipe'), as well as a smaller amount of 'Technical support', 'Advice', 'Self-help', where 'Advice' and 'Self-help' belong to the general register of 'Opinion', while the instructional texts belong to the general register of 'Non-opinion' (Egbert and Biber, 2013). The degree of recommendation is represented by variation in the A17 FTD.

The A8 (news) cluster combines the narrative registers which report an event ('News report/blog', 'Sports report'), as well as a smaller amount of 'Magazine article' and 'Obituary'. The presence of other registers in this cluster might be due to errors the annotators made in assigning these registers. 'Magazine article' is a very general category, which can contain more or less of argumentation (A1), entertainment (A5), information (A8) or evaluation (A17).

The A16 (definitions) cluster includes texts annotated with descriptive registers ('Encyclopedia article', 'Legal terms and conditions', 'Description of a person', a smaller amount of 'Description of a thing', 'Research article', 'Abstract'), as well as one narrative register ('Biographical story/history'). It is necessary to emphasise that all the texts annotated as belonging to these registers and to the A16 cluster also have a property of defining a topic, e.g., life of Peter the Great, a kind of cheese, a national holiday in India. We suppose that from the point of view of the annotators the distinction between biographical, descriptive and historical texts is less important. In this cluster a small number of texts (mostly 'Encyclopedia article') also have the secondary dimension A15 (text requiring specialist knowledge).

The A14A15 cluster contains texts belonging to the field of Science and Technology ('Encyclopedia article', 'Research article'), along with a small amount of 'Technical support', 'Technical report', 'How-to' and 'Instructions'), at this stage of experiment all texts of this cluster require readers to have background knowledge. The texts annotated as 'How-to' and 'Instructions' are technical, e.g. texts about how to write a compiler. The secondary dimension A16 in this cluster is also prominent. Our clustering procedure probably separated sci vs non-sci texts of the descriptive kind. Sci texts also tend to require more specialist knowledge in comparison to biographies.

The A1A12A17 cluster combines advertisements ('Description with intent to sell', 'Persuasive article or essay', 'Opinion blog', 'Review'). The A1A13A17 cluster mainly includes the texts which correspond to the Biber's 'Persuasive article or essay', but also a small number of other registers with persuasion and argumentation ('Prayer', 'Religious blog/sermon', 'Research article').

Even though the results of clustering are described in terms of the FTDs, the number of possible clusters will be much more extensive, when we include more texts in the analysis. Many existing clusters are better to be described with a set of FTDs, such as the A1A12A17 or the A1A13A17 clusters.

Having the preliminary results we suppose that some of our clusters are similar to Biber's general registers, e.g., the A7 cluster corresponds to Instructional texts, the A1A12A17 cluster to 'Intent to sell' and the A1A13A17 cluster to 'Persuasive articles, although there are some differences, e.g. we do not have lyrical or opinion clusters in this corpus. However, this is primarily because of the composition of our corpus.

'Persuasive article' and 'Review' are widespread across our clusters. The texts annotated as 'Persuasive article' are mostly presented in the A1A12A17 cluster and in the A1A13A17 cluster (adverts and argumentative propaganda texts), which is reasonable. There are only isolated cases of this register in the A7 cluster and the A8 cluster. Similar situation is with 'Review', which mostly occurs in the A1A12A17 cluster and sporadically occurs in the A8 cluster and the A16 cluster. We can expect a lot of variation in terms of the functional dimensions in texts of such registers, and this leads to the lack of internal stability of these registers in terms of the clusters they have been assigned to.

6. Conclusions

We have tested the Functional Text Dimensions framework on a wide variety of Russian texts and suggested changes to these dimensions in comparison to previous studies. For example, we removed largely duplicate dimensions with strong pairwise correlation and suggested new dimensions to make distinctions between important text types (see Appendix 1). Finally, we analysed the reliability of FTD annotation and obtained data to improve the annotation quality. The results of annotation are available from: http://corpus.leeds.ac.uk/serge/webgenres/

We have also experimented with various clustering techniques and detected patterns in annotated data, which lead to the possibility of uncovering "syndromes" of features corresponding to look'n'feel genres, such as 'News', 'Legal texts', 'Research papers'. However, identification of the clusters defined by several principal FTDs requires further research. In the initial steps for this research, we compared these clusters to other genre classification frameworks, in particular to registers from (Egbert and Biber, 2013). We detected cases of good agreement, as well as disagreement between the two annotation approaches, which can potentially lead to enrichment of both methods. In particular, some registers, e.g., 'Persuasive articles', exhibit internal variation, which is best explained by using the FTDs.

One of the important outcomes of the annotation experiment is that it demonstrates the possibility to achieve acceptable interannotator agreement on the FTDs, while the annotators often disagree with respect to atomic labels Sharoff et al. (2010). 'News' as a cluster corresponds nicely to the A8 principal dimension. However, this does not make 'News' a reliable label. On the contrary, a text with a high A8 value can be more or less argumentative (A1), light-hearted (A5), or can contain an overview of a topic (A16). A traditional genre palette forces the annotators to choose one label, which can be 'News report', 'Short story', 'Magazine article', 'Opinion', 'Persuasive article', etc. It is natural for different annotators to consider a text from different viewpoints, thus reducing reliability of their annotation. If linguistically similar texts

receive different labels, this in turn reduces accuracy of an automatic classifier. At the same, a combination of several FTDs is more likely to achieve both reliable annotation and reliable classification.

In the next step, we would like to train classifiers for each of these dimensions and apply them to a large Russian webcorpus of about 50 billion words Piperski et al. (2013). This should provide us with a genre map of the entire space of Russian web texts, so that the linguistic researchers can select a corpus subset according to their interests, e.g., personal narratives (A11) or evaluative texts (A17). To develop the classifiers we will need more research into the linguistic features associated with particular FTDs.

Acknowledgements

Research reported in this paper was partly funded by a grant provided by the Skolkovo Institute of Science and Technology. We are grateful to Alexander Piperski for his help in organising the evaluation procedure in the Russian State University for the Humanities and to Vladimir Selegey for general support.

References

1. *Adamzik K.* Textsorten—Texttypologie. Eine kommentierte Bibliographie. Münster : Nodus, 1995.
2. *Bocharov V., Bichineva S., Granovsky D.* et al. Quality assurance tools in the OpenCorpora project // Proc. Dialogue, Russian International Conference on Computational Linguistics. Bekasovo, 2011.
3. *Egbert J., Biber D.* Developing a user-based method of register classification // Proc. 8th Web as Corpus Workshop. Lancaster, 2013. July.
4. *Forsyth R., Sharoff S.* Document dissimilarity within and across languages: a benchmarking study // Literary and Linguistic Computing. 2014. Vol. 29. P. 6–22.
5. *Jakobson R.* Linguistics and poetics // Style in Language / Ed. by T. A. Sebeok. The M.I.T. Press, 1960. P. 350–377.
6. *Kilgarriff A.* The web as corpus // Proc. of Corpus Linguistics 2001. Lancaster, 2001. URL: http://www.itri.bton.ac.uk/techreports/ITRI-01-14.abs.html.
7. *Krippendorff K.* Reliability in content analysis: Some common misconceptions and recommendations // Human Communication Research. 2004. Vol. 30, no. 3. URL: http://faculty.washington.edu/jwilker/559/PAP/krippendorf-reliability.pdf.
8. *Lee D.* Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle // Language Learning and Technology. 2001. Vol. 5, no. 3. P. 37–72. URL: http://llt.msu.edu/vol5num3/pdf/lee.pdf.
9. *Piperski A., Belikov V., Kopylov N.* et al, Big and diverse is beautiful: A large corpus of Russian to study linguistic variation // Proc. Web as Corpus Workshop (WAC-8). 2013.

10. *Rehm G., Santini M., Mehler A.* et al. Towards a reference corpus of web genres for the evaluation of genre identification systems // Proceedings of the Sixth Language Resources and Evaluation Conference, LREC 2008. Marrakech, 2008.
11. *Rousseeuw P. J.* Silhouettes: a graphical aid to the interpretation and validation of cluster analysis // Journal of computational and applied mathematics. 1987. Vol. 20. P. 53–65.
12. *Salvador S., Chan P.* Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms // Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on. 2004. P. 576–584.
13. *Santini M., Mehler A., Sharoff S.* Riding the rough waves of genre on the web // Genres on the Web: Computational Models and Empirical Studies / Ed. by Alexander Mehler, Serge Sharoff, Marina Santini. Berlin/New York : Springer, 2010.
14. *Sharoff S.* Methods and tools for development of the Russian Reference Corpus // Corpus Linguistics Around the World / Ed. by D. Archer, A. Wilson, P. Rayson. Amsterdam : Rodopi, 2005. P. 167–180.
15. *Sharoff S.* In the garden and in the jungle: Comparing genres in the BNC and Internet // Genres on the Web: Computational Models and Empirical Studies / Ed. by Alexander Mehler, Serge Sharoff, Marina Santini. Berlin/New York : Springer, 2010. P. 149–166.
16. *Sharoff S., Wu Z., Markert K.* The Web library of Babel: evaluating genre collections // Proc. of the Seventh Language Resources and Evaluation Conference, LREC 2010. Malta, 2010. URL: http://corpus.leeds.ac.uk/serge/publications/lrec2010.pdf.
17. *Sinclair J.* Corpora for lexicography // A Practical Guide to Lexicography / Ed. by P. van Sterkenberg. Amsterdam : Benjamins, 2003. P. 167–178.

Appendix 1. Functional Text Dimensions

Code	Label	Question to be answered
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, if argumentation is obvious)
A3.	emotive	To what extent is the text concerned with expressing feelings or emotions? (None for neutral explanations, descriptions and/or reportage.)
A4.	fictive	To what extent is the text's content fictional? (None if you judge it to be factual/informative.)
A5.	flippant	To what extent is the text light-hearted, i.e. aimed mainly at amusing or entertaining the reader? (None if it appears earnest or serious; even when it tries to keep the reader interested and involved)
A6.	informal	To what extent is the text's content written in an informal style, using colloquialism and/or slang (as opposed to the "standard" or "prestige" variety of language)?

Code	Label	Question to be answered
A7.	instructive	To what extent does the text aim at teaching the reader how to do something? (e.g. a tutorial)
A8.	news	To what extent does the text appear to be a news report such as might be found in a newspaper, i.e. an informative report of recent events? (recent at the time of writing. None if a news source does not provide new information about what happened, while analysing information from other sources).
A9.	legal	To what extent does the text lay down a contract or specify a set of regulations? (This includes copyright notices.)
A11.	personal	Does the text report a first-person point of view?
A12.	compuff	To what extent does the text promote a commercial product or service?
A13.	ideopuff	To what extent is the text intended to promote a political movement, party, religious faith or other non-commercial cause? (i.e. any promotion of not-for-profit causes)
A14.	scitech	To what extent would you consider the text as belonging in the field of Science, Technology and/or Engineering? (As opposed to the Arts, Humanities &/or Social Studies. This is not necessarily a research paper. A newswire text can include scientific contents, so it can be judged as Strongly or Partly.)
A15.	specialist	To what extent does the text require background knowledge or access to a reference source of a specialised subject area in order to be comprehensible? (such as wouldn't be expected of the so-called "general reader")
A16.	encyc	To what extent does the text provide information to define a topic? (For example, encyclopedic articles or text books).
A17.	eval	To what extent do you judge the text to evaluate something? (For example, by providing a product review).
A18.	dialogue	To what extent does the text contain active interaction between several participants? (For example, forums or dialogue lines coming from theatre plays).
A19.	poetic	To what extent does the author of the text pay attention to its aesthetic appearance? ('Strongly' for poetry, language experiments, uses of language for art purposes)

Rating	Levels:
0	none or hardly at all;
0.5	slightly;
1	somewhat or partly;
2	strongly or very much so.

A PRODUCTION SYSTEM FOR INFORMATION EXTRACTION BASED ON COMPLETE SYNTACTIC-SEMANTIC ANALYSIS

Starostin A. S. (astarostin@abbyy.com),
Smurov I. M. (ismurov@abbyy.com),
Stepanova M. E. (mstepanova@abbyy.com)

ABBYY, Moscow, Russia

The article presents a mechanism for information extraction from unstructured natural language data. The key feature of this mechanism is that it relies on deep syntactic and semantic analysis of the text. The system takes a collection of syntactic-semantic dependency trees as input and, after processing them, outputs an RDF graph consistent with certain domain ontology.

The mechanism was implemented within a deployable information extraction system, which is a part of ABBYY Compreno technology—a powerful tool for a broad range of NLP-tasks that include machine translation, semantic search and text categorization. The description of the extraction algorithm and the results of the system performance evaluation are given.

Evaluation tests were conducted on the MUC-6 corpus. The overall F-measure we achieved using Compreno technology was 0.83, which is lower than the best results claimed by the researchers using machine learning approaches. Our system is still under development at the moment and we hope to improve its performance in the future. One of the advantages of Compreno technology is that, unlike many statistical approaches, it does not show an abrupt performance drop if the test corpus is changed. Thus Compreno demonstrates little dependence on the exact textual data it receives and therefore might be seen as a more universal and less domain-dependent solution. Our tests on the CoNLL corpus yielded an F-measure of 0.75 with no prior adjustments made.

Key words: information extraction, named entity recognition, syntactic analysis, anaphora and coreference resolution, production rule systems

Introduction

The article describes an information extraction method which is the core of the data mining system that has been in development by ABBYY over the last three years. This system is an integral part of a more universal text analysis technology known as ABBYY Compreno. Its key feature is the ability to perform complete syntactic-semantic analysis of the input text.

At the first stage a given input is analyzed by the Compreno parser [1]. The result is a collection of syntactic-semantic dependency-based parse trees (one tree per sentence). Nodes and edges of each tree are augmented with diverse grammatical and semantic information. The parse tree forest is then used as input for a production system of information extraction rules. The application of the rules results in the formation of an RDF graph consistent with a domain ontology.

In the first section of the article we provide a description of the information extraction mechanism. We briefly describe the input data, the method used to store extracted information, the structure of the extraction rules and the algorithm of their implementation.

The approach we propose demonstrates two significant advantages. Firstly, the availability of syntactic and semantic structure allows us to extract facts as well as entities. Fact extraction rules that rely on the structure of syntactic-semantic trees tend to be laconic yet highly efficient, easily covering most natural language expressions. Secondly, the system shows little dependence on a particular language. Since our parse trees contain language-independent data (like semantic roles or universal semantic classes), many extraction rules are universal and can be used for different languages.

Despite the fact that we use declarative rules in our system, our approach to information extraction cannot be described as a rule-based one, because the syntactic and semantic analysis that precedes the extraction is not based on a set of rules. The sort of analysis performed by the Compreno parser can be defined as model-based: it rests upon a multilevel model of natural language created by linguists and then corpus-trained. Thus it is possible to consider our method hybrid, it being model-based at the first (preparatory) stage and rule-based at the second.

In the second part of the article we provide the results of the tests we conducted to evaluate our system's performance. We used the MUC-6 corpus to run the tests and chose a standard set of information objects (Person, Organization, Location and Time) for evaluation.

Information Extraction Mechanism

The input accepted by the information extraction mechanism is a sequence of syntactic-semantic trees (one tree per sentence). These trees are generated by the Compreno parser during the analysis. Each tree is projective and its nodes in most cases correspond to the words of the respective sentence, although there are some null-nodes with no surface realization. Nodes and edges of a tree are augmented with grammatical and semantic information. More details on the input and the Compreno parser can be found in the full version of the paper available on the conference website, or in [10].

The output of the extraction mechanism is an RDF graph. The idea of RDF (Resource Definition Framework, [9]) is to assign each individual information object a unique identifier and store the information about it in the form of SPO triples. S stands for subject and contains the identifier of an object, P stands for predicate and identifies some property of an object, O stands for object and stores the value

of that property. This value can be either a primitive data type (string, number, Boolean value) or an identifier of another object.

All the RDF data is consistent with an OWL-DL[1] ontology [7] which is predefined and static. Information about situations and events is modelled in a way that is ideologically similar to that proposed by the W3C consortium for defining N-ary relations [2]. The consistency of the extracted information with the domain model 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 entities to the respective parts of the original text. The combination of an RDF graph and annotation links will hereinafter be called an annotated RDF graph.

An annotated RDF graph is generated at the very final stage of the information extraction process. Until that we use a more complex structure to store information during the process. This structure can be described as a set of noncontradictory statements about information objects and their properties. Further on we will often call that a "bag of statements". Running a few steps forward, we have to note that all the statements are generated during the process of information extraction rules' application.

A bag of statements has several important properties:

1. **Cumulativity**. Statements can be added to but not removed from a bag.
2. **Consistency**. All the statements in a bag are non-contradictory to each other.
3. **Consistency with ontology**. A bag of statement can anytime be converted into an annotated RDF graph consistent with certain ontology.
4. **Transactionality**. Statements are added in groups, and if any statement of a group contradicts other statements from the bag, the addition of the whole group is cancelled.

The final annotated RDF graph can also be viewed as a bag of statements, if each SPO triple and each link from an object to a segment of text is considered a statement about that object. Therefore it is important to point out the difference between our temporary information storage structure (the inner structure) and the final output in the form of an RDF graph.

The main distinction is that the statements from the inner structure can be used to create functional dependencies, i.e. some statements may depend on the presence of others. For instance, we can state that a set of values of a certain object`s property should always contain a set of values of some other property of a different object. If the set of values of the second object is changed, the first object's property changes as well. We will hereinafter refer to such statements (which use functional dependencies) as *dynamic* statements. Another difference of the inner structure is that it may contain some auxiliary statements that do not comply with the final annotated RDF graph structure and are used only during the extraction process.

---

[1] The OWL DL language subset that we use is similar to OWL Lite, but we also exploit DisjointWith axiom.

Here is the list of the possible statement types:

1. **Existence statement**. A statement that proclaims the existence of an information object and creates unique identifiers for them.
2. **Class membership statement**. A statement that attributes an object to a certain class in the ontology.
3. **Property statement**. A statement that defines some property of an object.
4. **Annotation statement**. A statement that connects information objects to parts of the original input text.
5. **Anchor statement**. A statement that links information objects to parse tree nodes, which enables us to access these objects again during the extraction process.
6. **Identification statement**. A statement that merges objects which refer to a single real-life entity.
7. **Functional restriction**. A function, returning a Boolean value, which makes it possible to impose additional restrictions on certain groups of objects. After a function has been added to a bag of statements no statement that would make the function false can enter the bag.

Figure 1 contains schematic diagrams of all statements types available in our system. One can see that only statements of four types may be dynamic. Identification, anchor and existence statements may not depend on other statements.

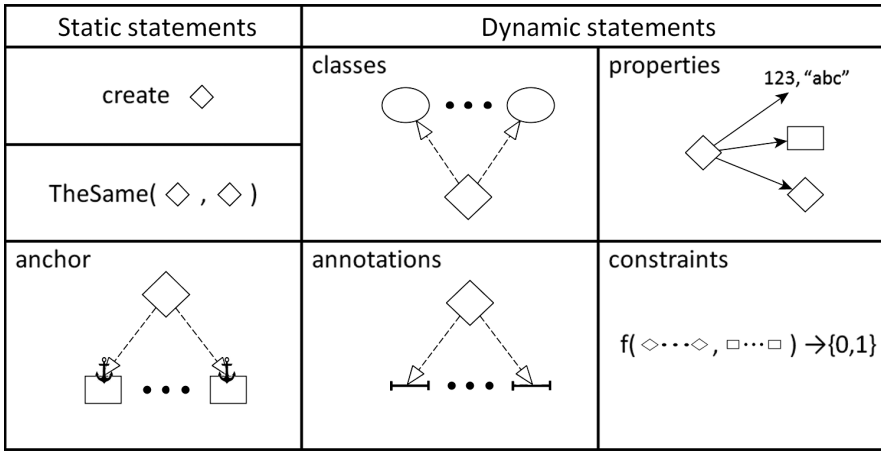


Fig. 1. Types of statements used in the information extraction process. Diamonds represent information objects (individuals), ellipses represent classes (or concepts) and rectangular boxes represent parse tree nodes

Let us describe anchor statements more thoroughly because they are a very important part of information extraction mechanism. Anchor statements link information objects to parse tree nodes, which enables us to access these objects continuously during the extraction process. The term 'anchor' was coined when the system was

in development so that the links between objects and tree nodes could be easily referred to. One object can be anchored to a set of nodes via a number of anchor statements.

The interpreter of the information extraction rules handles these anchors in a special way: the left-hand side (or condition side) of a rule in our system can contain so-called object conditions. These conditions require object(s) with certain properties to be anchored to a node before the rule can be executed. And if the object was found and the production executed, this object can be accessed in the right-hand side of the rule.

Object conditions are most widely used in the rules that extract facts, but they are quite useful with named entities as well, since they make it possible to break the extraction process down to several simple stages. For instance, one rule might only create an unspecified Person entity, while the following ones add properties like first name, surname, middle name and alike. It has also become quite common to create auxiliary objects which serve as dynamic labels of parse tree nodes. First some rules create these auxiliary objects and attach them to certain nodes, and then other rules check for these objects with help of object conditions in their left-hand sides.

Detailed information about other types of statements can be found in the full version of the article, available on the website of the conference.

Information Extraction Rules

Information extraction process is controlled by the production rule system. There are two types of rules in the system: parse subtree interpretation rules (or simply interpretation rules) and identification rules. Since interpretation rules are much more frequent, whenever we do not specify the exact type of rule the reader should assume that it is an interpretation one. Information on both types of rules can be found in the full version of the article, available on the conference website.

During the development of the extraction mechanism several goals were pursued. In the first place, our intention was to exploit such advantages of the production rule systems as modularity [8] and separation of knowledge from the procedure. We particularly wanted to relieve the developers from the necessity to order the rules[2]. Secondly, we intended to implement an efficient deterministic output model. Speaking in terms of traditional production systems [3] we can define parse tree forest and a bag of statement as our knowledge base, while the extraction process itself can be described as a forward chaining inference process.

---

[2] One particular example of quasi-production language that does not comply with this requirement is Jape [5]. Jape requires setting the order in which groups of production rules (or phrases) are executed explicitly. During their execution rules within a group do not have the access to each other's results. In the process of development of such rules it often occurs that the rules which create an object of a certain type are executed after the rules which accept such an object as their input. Moreover, it is not possible to reorder the rules because rules from the first group might also use some objects created by the second group. The only solution to this problem is to launch the same groups of rules several times. However, this solution is far from ultimate since it artificially limits the number of recursion steps.

Information Extraction Algorithm

While describing the information extraction algorithm we use the generic term 'matching'. By this term we mean both matching of a tree template in an interpretation rule with a subtree of an actual parse tree and matching of an identification rule with a certain pair of objects. Formal definition of matching can be found in the full version of the article, available on the conference website. Here we will just point out that finding a matching is a sufficient condition for the right-hand side of the rule to be converted into a set of statements.

The information extraction algorithm has the following steps:

1. Analyze the input text with the Compreno parser to get a forest of syntactic-semantic parse trees.
2. Find all the matchings for the interpretation rules that do not have object conditions.
3. Add the matching to the sorted matching queue
4. If the queue is empty, terminate the process.
5. Get the highest priority matching from the queue
6. Convert the right-hand side of the production a group of statements.
7. Try to add the statements to the bag.
8. If failed, declare the group of statements invalid and go to step 4.
9. Else if succeeded, look for new matchings.
10. If found, add new matches to the queue Go to step 4.

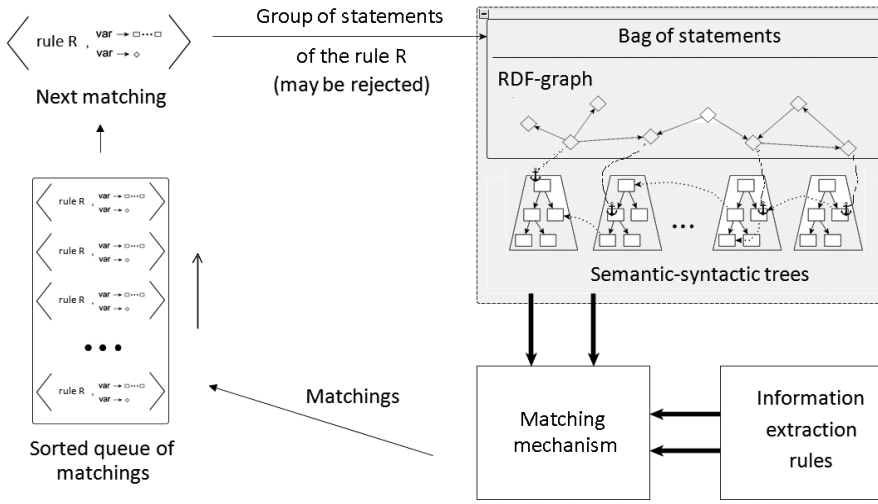


Fig. 2. Schematic representation of the information extraction process

Some parts of the above algorithm need to be described more thoroughly. Steps 2 and 9 are performed with the help of a special matching mechanism. This mechanism can retrieve all the matches for the rules without object conditions. It also constantly monitors the contents of the bag of statements. Every time step 7 is performed successfully and new statements get into the bag, the mechanism takes them into account and, if necessary, generates new matches for the rules that do contain object conditions. These new matchings can be created both for the rules that have already been matched before and for those which remained unmatched until that moment. The former occurs when an object condition of a certain rule is matched by more than one object. In this case each object is matched in a separate matching.

The implementation of the matching mechanism is relatively complex. For one, it has a built-in bytecode interpretor for the compiled rules, a system of indexes for the syntactic-semantic trees, a module for tracking changes in the bag of statements and several other features. Full-length description of this mechanism is beyond the scope of the paper.

It is also important to explain the way the queue of matchings is sorted at the third step. In some cases developers can set the order of rules, i.e. there is partial order over the whole set of rules. Of any two rules one can be given priority over the other. It means that if both rules are ready to be executed, the rule with the higher priority should go first. For convenience reasons we also support group ordering of rules. If group A was given priority over group B, then each rule belonging to group A has higher priority than one belonging to group B. Partial order relation is transitive. Correctness of partial order is checked every time a system of rules is compiled. If loops are detected, compilation fails and the user receives an error message. The order of matching in the queue is always consistent with the partial order set within a system of rules. This approach differs significantly from those with consecutive execution of rules, since partial order only determines the priority of rules and does not prevent repeated execution.

It is easy to see that the described algorithm does not consider alternatives. If some matching is inconsistent with the bag of statements in its current state, it is simply dismissed. We can afford to use this 'greedy' principle because our parser performs word-sense disambiguation, so we rarely ever have to hypothesize about a node. There some exceptions like words unknown to the parser, but for such cases we have special methods of dealing with these words and incorporating them in our model.

Evaluation

We tested our system on the texts that were manually annotated with name entities for the 6th Message Understanding Conference (MUC-6) held in November 1995 [4]. Today the MUC-6 data set is considered one of the main evaluation benchmarks for named entity recognition. You can find the detailed description of the evaluation process in the full version of the article. In the paper version we limit ourselves only to results, which are shown in the table below:

Table 1. Evaluation results

Type of entity	Precision	Recall	F-measure
All entities	0.853	0.813	0.832
Money	0.947	0.933	0.940
Person	0.700	0.887	0.783
Location	0.936	0.806	0.866
Organization	0.767	0.639	0.697
Date	0.941	0.880	0.910
Time	0.674	0.573	0.620

These results are lower than the numbers shown by machine-learning systems on MUC-6 original test corpus of 30 texts (the F-measures of many systems that participated in the contest were higher than 90% and the winner reached 96,42% in F-measure). However it is worth noting that our system was not specifically trained on MUC-6 corpus texts or any other WSJ articles. We also did not make any deliberate changes in our model (apart from the technical ones, see the description in the full version of the article) that could artificially improve performance on this particular set of texts. It would be correct to assume that our system was put in position of a statistical entity extractor trained on a completely different corpus.

Error analysis demonstrated that approximately 60% of the errors were the errors of the Compreno parser, 20% occurred due to flaws in the extraction rules and the MUC-6 corpus inconsistencies accounted for the remaining 20%. These results show that the system has a significant potential for further development, especially since there are hopes to improve the quality of the syntactic-semantic parser.

After testing our system on MUC-6 corpus we also conducted additional tests on CoNLL corpus [6]. During these tests no settings were modified and no changes were made whatsoever. The resulting F-measure was 0,75. This allows us to make a preliminary conclusion that our system is more resistant to the replacement of one corpus with another than systems based on machine-learning approaches. In the near future we intend to conduct a more extensive performance evaluation on several other corpora.

We do realize that the tests we conducted are insufficient to provide complete evaluation of the system performance (and give the reader full insight of the system), especially since the spectrum of its applications is much wider than named entity recognition.

Conclusion

In this paper we described an information extraction mechanism based on a production rule system. The rules are applied to the results of full syntactic-semantic analysis performed by the Compreno parser. The output of the extraction mechanism is an RDF graph consistent with domain ontology and augmented with information about annotations of extracted individuals.

We also presented the idea of storing extracted information as a set of dynamic logical statements. We mentioned two types of declarative extraction rules: interpretation rules that interpret subtrees of syntactic-semantic trees and identification rules

that merge information objects. We gave schematic description of the information extraction algorithm.

A considerable advantage of the system we have created is that a developer of rules does not have to set the order of their execution. Rules are executed in arbitrary order if there is data that matches their left-hand sides. However, if the necessity appears, the developer can set partial rule order.

Finally, we present the results of the evaluation tests we conducted on the MUC-6 manually annotated corpus. Our system demonstrated relatively good performance with no prior adjustments made. Additional tests on the CoNLL corpus allow us to make a preliminary conclusion that our system is not dependent on a particular corpus (like statistical ones often are) and remains efficient after the corpus is changed. To confirm this conclusion further tests are required and we plan to conduct them in the nearest future. After these tests are performed we intend to publish a new article focusing on the task of fact extraction.

References

1. *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.
2. Defining N-ary Relations on the Semantic Web, available at http://www.w3.org/TR/swbp-n-aryRelations
3. *Gavrilova T. A., Khoroshevskij V. F.* (2000) Knowledge Bases of Intellectual Systems [Bazy znanij intellektual'nyh system], Piter, St. Petersburg, Russia
4. *Grishman R., Sundheim B.* (1996), Message Understanding Conference—6: A Brief History, available at: http://acl.ldc.upenn.edu/C/C96/C96-1079.pdf.
5. *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, Amsterdam.Language-Independent Named Entity Recognition, available at http://www.cnts.ua.ac.be/conll2003/ner/
6. OWL Web Ontology Language Overview, available at http://www.w3.org/TR/2004/REC-owl-features-20040210
7. *Pospelov D. A.* (1989) Modelling Reasoning. Experience in the Analysis of Mental Acts[Modelirovanije Rassuzhdenij. Opyt Analiza Myslitel'nyh Aktov]. Radio i Svayz, Moscow, Russia.
8. Resource Description Framework, available at http://www.w3.org/RDF/
9. *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.

ОПЫТ СОЗДАНИЯ СИНТАКСИЧЕСКОГО АНАЛИЗАТОРА АРАБСКОГО ЯЗЫКА ДЛЯ ПРОМЫШЛЕННОГО ПРИМЕНЕНИЯ

Стребков Д. Ю. (strebkov@dictum.ru),
Хилал Н. Р. (hilal@dictum.ru),
Руджеймийя А. (redjaimia@dictum.ru),
Скатов Д. С. (ds@dictum.ru)

ООО «Диктум», Нижний Новгород, Россия

Ключевые слова: синтаксический анализ, синтаксический анализатор, семитские языки, арабский язык

THE EXPERIENCE OF BUILDING INDUSTRIAL-STRENGTH PARSER FOR ARABIC

Strebkov D. Y. (strebkov@dictum.ru),
Hilal N. R. (hilal@dictum.ru),
Redjaimia A. (redjaimia@dictum.ru),
Skatov D. S. (ds@dictum.ru)

Dictum Ltd., Nizhny Novgorod, Russia

We present a propagation of a hybrid approach for natural language parsing on Semitic languages on the example of the Arabic language. The hybrid approach proposes a way for acquiring dependency and constituency parses simultaneously at every step of the analysis. The result of the propagation is represented by a syntactic parser for Arabic language and the fact that the parser shows quite satisfactory results and belongs to the group of rule-based parsers actually forms scientific novelty of this article. We give a short review of Arabic Natural Language Processing (NLP) technologies and their current state and then describe steps that were required for our propagation: choosing of morphological analyzer, morphological index compression scheme, description of rule base system that is used by the parser, modifications that were needed for tuning in the core parsing algorithm. We also designate problems that we faced during the propagation and the results that we finally achieved. In the end we provide results of brief evaluation of the parser and give information on its current usage.

Keywords: syntax parsing, syntax parser, Semitic languages, Arabic language

1. Introduction

Arabic, which is the mother tongue of more than 300 million people, has received substantial attention by modern computational linguistics basing on its morphology and flexible sentences construction. The scale of Arabic-related research work is now orders of magnitude beyond what was available a decade ago [10]. At the same time, the language presents significant challenges to many natural language parsing applications for several reasons. Arabic sentences are syntactically ambiguous and complex due to the frequent usage of grammatical relations, order of words and phrases, conjunctions, and other constructions such as diacritics (vowels), which are known in written Arabic as "altashkiil" [1].

Result of the above interest can be presented as several applications. Their main goals are parsing Arabic language and providing some helper features for that. In this article we briefly describe some of such applications. Among them are three syntactic parsers (Stanford Parser [9], Berkeley Parser [16] and LFG Rule-basic Parser [2]), two morphological analyzers (Buckwalter Morphological Analyzer [7] and ElixirFM [18]), and Part-of-speech tagger (POS tagger) application developed by Stanford NLP Group [23].

Having these Arabic NLP applications available, the most significant motivation for the development of another parser are the existing NLP modules that we have developed and which are used industrially:

- Dictum's syntactic parser is based on the hybrid approach for NLP and has language-independent core component designed to support right-to-left (RTL) languages as well;
- "key-value" model for compact store of linguistic information that supports efficient access;
- opinion mining application which also has language-independent core and has been developed to deal with syntactic parser results presented in a specified format.

These applications are designed to be flexible for tuning and extending. Finally our model for syntactic rules representation supports semantic information marks.

2. Linguistic resources

In this paragraph we give a brief description of the existing syntactic modules for Arabic.

2.1. Syntax Parsers

It is necessary to mention that most accurate Arabic parsers are based on data-driven approach and assume using treebanks to learn probabilistic context-free grammars (PCFG) which assign a sequence of words the most likely parse tree [9]. Among them are Stanford Parser and Berkeley parser.

Stanford Parser is a statistical parser created by Stanford Natural Language Processing Group. Used to parse input data written in several languages such as English, German, Arabic and Chinese, it has been developed and maintained since 2002. The Arabic component takes the text as input and returns part-of-speech tagged text (the parser uses Stanford POS tagger for that) and a context-free phrase structure grammar representation:

```
Your query: هذا الرجل هو سعيد .
Tagging: هذا/DT الرجل/DTNN هو/PRP سعيد/NNP ./PUNC
Parse:
(ROOT
  (FRAG
    (NP
      (NP (DT هذا))
      (NP (DTNN الرجل)))
    (NP (PRP هو))
    (NP (NNP سعيد))
    (PUNC .)))
```

Fig. 1. Example of Stanford parser's phrase structure grammar representation

Arabic version of Stanford parser is based on the Penn Arabic Treebank (PATB) and uses phrasal category set of it [15]. Also the parser assumes precisely the tokenization of Arabic used in the PATB. There is no grammatical relations analysis available for Arabic. As for performance of Stanford parser, the dependency accuracy of the parser is around 83.5%.

Berkeley Parser is The Berkeley Natural Language Processing Group's parser; it is based on PCFG as well. Just like the Stanford parser, it returns a phrase structure representation of the input text in terms of PATB phrasal category set. Berkeley's PCFG is created using split-and-merge training strategy: splitting provides a tight fit to the training data, while merging improves generalization and controls grammar size. The resulting grammar is remarkably good at parsing [16]. According to the [9], that parser shows most state-of-the-art performance and leaves Stanford Parser behind: the accuracy of the Berkeley's parser is around 84%.

In addition to PCFG based parsers there is a rule-based parser for Arabic language, and it is the only one to our knowledge.

Arabic LFG Rule-basic Parser is the first Arabic rule-based parser available for Modern Standard Arabic (MSA). It was implemented using Xerox Linguistics Environment (XLE). Since the parser is based on LFG grammar [2], its output is represented by its special structures as shown on example below:

Your query : على الطريق رجل
Parse:

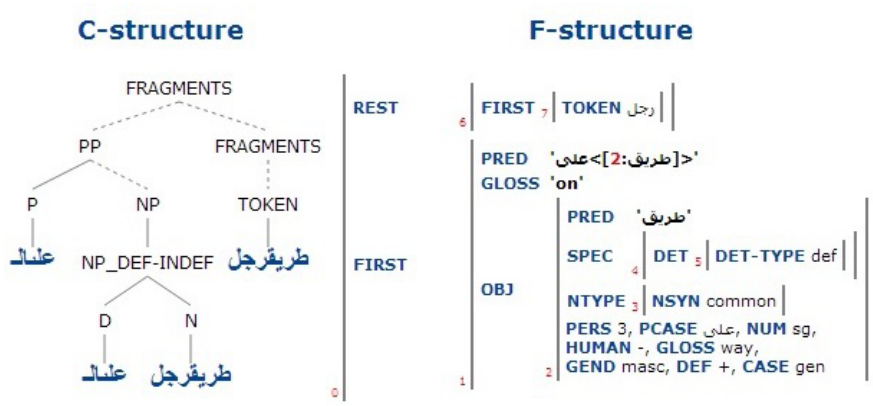


Fig. 2. LFG Rule-basic Parser's result

Figure 2 shows us parsing result in terms of phrase structure (c-structure) and grammar attribute-value pairs (f-structure).

According to the evaluation results, the accuracy of the parser is around 87% [2]. At the same time it is necessary to highlight that the result was got on a rather small corpus that consisted of 69 manually collected sentences only. M. Attia also noticed that he concentrated on short sentences and used robustness techniques to increase the coverage. All of these use hand-crafted grammars, which are difficult to scale to unrestricted data [22].

There are also Bikel Parser [12] and Malt Parser [13] which also belong to the group of data-driven parsers, so that approach is the most popular in case of Arabic language parsing.

2.2. Morphology

Morphological ambiguity in Arabic is an acute problem due to the richness and complexity of Arabic morphology.

The deficiencies of Buckwalter Morphology. Despite the fact that Buckwalter Morphology is a stem-based database and has been considered as the "most respected lexical resource", it includes a large number of entities which are not used in contemporary Arabic texts and this fact reduces the benefit of Buckwalter Morphology in analyzing the modern language. In addition, Buckwalter has some significant problems [2]:

- Absence of imperative state of almost every verb.
- Not all verbs have their correct passive form in correct tense.
- Large number of obsolete words.
- Misspelled words which lead to a massive increase in the ambiguity level for correct words.

For the reasons below, we decided to use the Elixir FM program for generating our own morphology instead of Buckwalter's:

- The lexicon's format considers the diacritics, and it means that for each of the entities ElixirFM program sets the correct vowel marks.
- In ElixirFM each verb has its correct passive form, tense and state.
- ElixirFM does orthographic analysis to get correct grammar meaning for (two, three-token) entry. For example, the word (فهم) can be interpreted as one-token "bare entity" or as tow-token "entity involved conjunction". ElixirFM includes both of these two variants in our morphology.
- ElixirFM uses the features of both word segments and the root to determine the morpho-syntactic features of the input inflected word.

Fig. 3. ElixirFM output

Figure 3 shows us analysis results of a given word: stem, transcription, grammar values and vowel reconstruction.

New morphological groups as expansion of ElixirFM issuance. Despite all advantages of the ElixirFM, the set of grammatical meanings which it gives do not cover the whole syntax of the Arabic language. In order to fill this gap we had to expand the list of grammatical meanings and add groups invented by us such as Condition, Special Function Word, Emotional Interjection and Preference name.

Also, we had to correct errors in the output of ElixirFM connected with some functional and frequency words, that were the reason for fault in the previous syntactic analysis. For example, entry (و) had two homonyms with different grammatical meanings, the first and the correct one = Conjunction, and the second erroneous = Preposition.

In case of adverbs, most of them were mistakenly identified as adjectives. To fix this error we added a check for both the case and the last letter. If the adjective was in accusative case and ended with letter Alif "ا", it became automatically adverb.

ElixirFM does not give complete information about irregular genders and does not have genders for such nouns as Broken Plurals. We assembled in lists all Broken Plurals with their numbers and genders that resulted in a full actuation of the syntactical rules with successful checking both the gender and number of nouns and, therefore receive the correct parsing.

It is necessary to mention that ElixirFM does not provide any information about control models of Arabic verbs, so currently acquisition of that very valuable information is a plan for further extension of our system.

As it was mentioned in the Introduction part, we use an efficient "key-value" model for compact storing of linguistic information (DAWG) [19]. After choosing ElixirFM as a source of morphological information, the next task was to find a set of Arabic words that could be passed to ElixirFM. It would have provided required morphological information that could have been stored in a text file. That morphological dump actually is an intermediate representation of the parser's morphology component: after being generated, it could be modified later by adding new morphological groups. As for the source of linguistic data, finally we fix on a combination of these two resources:

- Arabic Wordlist for Spellchecking that contains 9 million words [5];
- Twitter archive. We extracted all unique words from it getting around 1 million words.

Morphology data storage problem. Having that morphological dump created, we use a morphology index generation program that stores linguistic information from the dump to the DAWG [19]. The subset of grammar value and normal form being stored gave us the serialized representation of 140 Mbytes, while having 2 Mbytes for English, and 8 for Russian, so the size needed to be fixed. The structure of Arabic morphological system could be presented as a combination of two layers [3]. The former, derivation layer, is non-concatentative and opaque in the sense that it is a sort of abstraction and does not have a direct explicit surface manifestation. The latter, inflection layer, applies concatentative process by using prefixes and suffixes to express morphological syntactic features. The derivation uses interdigitation—a process when Arabic words are formed through the amalgamation of two tiers, namely, a root and a template. A root is a sequence of three consonants, and a template is a pattern of vowels with slots into which the consonants of the root are inserted:

Table 1. Interdigitation example

Pattern	**$R_1aaR_2iR_3$**			
Root	KTB كتب	QTL قتل	FHM فهم	SRB شرب
Stem	KaaTiB كَاتِب	QaaTiL قَاتِل	FaaHiM فَاهِم	SaaRiB شَارِب

The example above shows how four different stems could be formed from one pattern ($R_1aaR_2iR_3$) using corresponding roots. As for the number of different patterns in the Arabic language, there are around 500 of them [4] and it is possible to get all stems for the root by applying to it all available patterns.

Taking into account the fact that DAWG is better compressed if the keys do have many common prefixes and suffixes [19], which is not true for Arabic by default, the

decision for optimization was to split each stem on two basic parts: `root` part and `pattern` part:

$$KaaTiB \rightarrow KTB,\ R_1aaR_2iR_3$$

Secondly, ElixirFM provides information from inflection layer, i.e. shows `prefixes` and `suffixes` that were used during word formation. Due to it, we can change the conception of the key in our morphology index: for each Arabic word we store it in the following format instead of storing it as it is:

pattern|prefix|suffix|root

The figure 4 shows the advantage in compactness of such keys representation in comparison with straightforward approach when Arabic words act as keys:

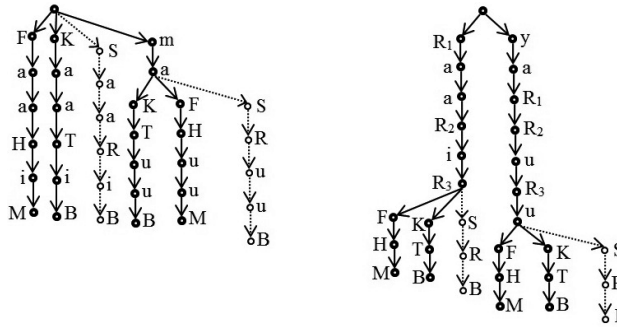


Fig. 4. Example of two approaches for keys representation

Both prefix trees are formed of four words: KaaTiB, FaaHiM, maKTuuB, maFHuuM. The left tree shows straightforward approach of keys representation, and the right one—approach that uses splitting technique mentioned above. As it could be seen from the figure 4, adding new roots (such as SRB) is more efficient from tree's size point of view.

Having splitted approach implemented, we reached the size of morphology index to be around 50 Mbytes.

3. Parser

In this chapter we mainly focus on modifications of the core of our language-independent syntactic parser that were required to get it working with Arabic language. The detailed description of the hybrid approach that was an inspiration of our parser is available in [20].

3.1. Rule base

As it was mentioned in the Annotation, our parser is a rule-based one. We used principles and approaches listed in our paper [20] to compile syntactic rules for Arabic, and at the moment their number is 193. These rules reflect and consider the specifics of the Arabic syntax. For example, in Arabic we can find, with the same frequency, (subject—verb) and (verb—subject) and this means the existence of two symmetric rules with the same priority. But with objects the (object—verb) version is more often than (verb—object) version, and therefore only one symmetric rule takes the priority which is determined by a check (!PH.InvertedLinksCount). If the rule doesn't take the priority we use check (PH.InvertedLinksCount), as mentioned in figure 5:

```
//التفاحة أكل   "apple eat"
Action+EntityObject {
  T: [Action] [Entity]
  C: LI1.Voice == VOICE_ACTIVE && ((LI2.Case == CASE_ACC &&
  !LI1.HasPersonalPron) || (LI2.Case == CASE_GEN && LI2.HasPrep)) &&
  PH2.Type != PHRASE_RELATIVE_PRON
  && PH2.Type != PHRASE_PERSONAL_PRON;
  Main: 1; L: 1=>VerbControl=>2; S: 1=>Object=>2;
  A: PH.InvertedLinksCount = 1;
}
```

Fig. 5. Description of "Action+EntityObject" rule

Arabic grammar has special categories for words that shift one or more elements of a clause into the accusative case. One of these categories is particle "Inna and her sisters" "ان واخواتها" which is usually used as subordinating conjunctions. It requires that the subject of the subordinate clause is in the accusative case and the predicate in the nominative case. In our morphology we have identified these particles in a separate group called "Special Function Word" and tried to describe it through our syntactic rules as follows:

```
//كأن المطر سيهطل - ليت الشباب يعود  "as if the rain come – wish
youth returned"
Action+Entity+SpecialFuncWord  {
  T: [Action] <> [Entity] <> [SpecialFuncWord]
  C: LI2.Case == CASE_ACC && PH1.Type != PHRASE_IMPERATIVE_ACTION
  && LI1.Gender == LI2.Gender;
  Main: 1; L: 1=>PredSubj=>2; 2=>Auxiliary=>3;
}
```

Fig. 6. Description of "Action+Entity+SpecialFuncWord" rule

Similarly, the category of verbs "Kana and its sisters" "كان و أخواتها" has the effect of shifting the predicate (خبر كان) from the nominative case to the accusative case. These verbs all denote existential states of being (or not being), becoming and remaining. We put these verbs in a group named Special Verbs and described it in syntactic rules.

Another special category of words and particles is the exclamation of wonder "اسلوب التعجب". It forms from particle (ما) and elative form which is identical with a verb form IV (af3ala). To identify the verb form IV in the whole morphological dictionary we put each adjective beginning with letter (j) in special group named Preference Name. Then, we described the exclamation of wonder in syntactic rule as follows:

```
//ما أجمله !   "how lovely it is !"
Wonder With (ما) {
  T: [PreferenceName] <> ["ما"]
  C: LI1.Case == CASE_ACC && PH2.Type == PHRASE_RELATIVE_PRON;
  Main: 1; L: 1=>Quantifier=>2;
  S: 1=>Quantifier=>2;
}
```

Fig. 7. Description of "WonderWith" rule

Also, vocative particles which come before noun are often used in Arabic and they can place noun into one of two cases (nominative or accusative). In our syntactic rules we described a vocative particle "yaa" (يا) and got a new phrase named "PHRASE_REQUIRED", that inherits the properties of one of the components PH1 or PH2 and can be used in other rule templates, as shown in figure 8 and figure 9:

```
// يا ولد   "Oh boy"
Entity+Call  {
  T: [Entity] <> ["يا"]
  C: LI1.Case != CASE_GEN && PH2.Type == PHRASE_EMOTIONAL_INTERJ;
  Main: 1; L: 1=>Auxiliary=>2;
  A: PH.Type = PHRASE_REQUIRED; LI.Gender == LI1.Gender && LI.Number
  == LI1.Number;
}
```

Fig. 8. Description of "Entity+Call" rule

```
//يا ولد ، اذهب    "Oh boy, go"
Any+Coma+Required {
  T : [Action] ({,} | {،}) [Required]
  C: LI1.Gender == LI2.Gender && LI1.Number == LI2.Number &&
  LI1.Person != PERSON_1ST;
  Main: 1; L: 1=>Parenth=>2;
  J: 1<=IsolEnd; 2<=IsolBegin;
}
```

Fig. 9. Description of "Any+Comma+Required" rule

Also, we devised a syntax rules that are able to analyze more complex structures such as the Subordinate Clause by using functions. In figure 10 we can see the relation between action and relative pronoun, which introduces a relative clause. As a result, we get entity equipped with specific function named ClauseEmbedded(PH) in the section A:

```
// الـذي نـجح    "который succeeded"
Action+Relative=Entity {
  T: [Action] <> [Relative]
  C: LI2.Gender == LI1.Gender && LI1.Person == PERSON_3RD;
  Main: 2; L: 2=>Subord=>1;
  A: PH.Type = PHRASE_ENTITY; ClauseEmbedded(PH);
}
When this function is needed we invoke it as IsClauseEmbedded(PH1)
in section C:
 //؟ من الـذي نـجح    "кто есть который succeeded ?"
Question+ActionWith(من)=Subject  {
  T: {؟}  [Entity] <> "من"
  C: IsClauseEmbedded(PH1) && PH2.Type == PHRASE_INTERROG_PART
  && LI1.Person != PERSON_1ST && !LI2.HasPrep;
  Main: 1; L: 1=>Quantifier=>2;
}
```

Fig. 10. Description of "Action+Relative=Entity" and "Question+ActionWith=Subject" rules

In general, our rule base covers all commonly used language constructs such as nominal and verbal sentence, compound sentences, conditional expressions.

3.2. The algorithm

The structure of the algorithm. Our parser uses an algorithm which can be treated as a combination of Cocke-Yanger-Kasami and Eisner's parsing algorithms ([17] and [8] correspondingly), to find dependency trees by corresponding phrase trees created by rules described above. The figure 11 shows an example of CYK's interpretation. As in usual implementation of the algorithm, it starts with upper-triangular matrix `M[2][2]` and fills its main diagonal with one-token phrases; each phrase from some cell is created from some grammatical value of corresponding token. Phrases $\texttt{Noun}_1$, $\texttt{Adj}_1$ and $\texttt{Noun}_2$ are created on that iteration.

An important moment here is the choice of destination cell for the phrase. Since tokens are numbered from right to left in case of the Arabic language, the algorithm creates phrase $\texttt{Noun}_1$ as the first one, $\texttt{Adj}_1$ and $\texttt{Noun}_2$ only after that. Therefore, if we will not take that fact into account, the phrase $\texttt{Noun}_1$ will be placed into the leftmost cell—`M[0][0]`, and that will look confusing because the phrase actually corresponds to the rightmost token of the sentence. To prevent that effect, we made a RTL-specific modification in the algorithm that adds phrases to the matrix `M` in reverse order, so phrase $\texttt{Noun}_1$ is placed into `M[1][1]` cell:

اللاعبون المخلصون

^{2}المخلصون : **Adj$_1$** ^{3}المخلصون : Noun$_2$	**Noun$_3$: Rule=Adj+Entity** Noun$_4$: Rule=Entity+Noun
	^{1}اللاعبون : **Noun$_1$**

Fig. 11. CYK matrix

That reverse filling process affects all diagonals of M, not only the main one.

On the next iteration the algorithm starts moving from the main diagonal to the next diagonal in upper-right direction, and we start creating phrases that cover 2 consecutive tokens. Just in that iteration we start using our rule base: from that moment for each cell that we are filling we iterate all rules from rules base, and if a rule passes template and criterion checks, we create a new phrase and add it to the cell that we are currently filling (phrases $Noun_3$ and $Noun_4$ are created that way). That iteration is the last one for our example; and phrases from upper-right edge cell cover entire input sentence.

4. Evaluation

The current evaluation of the described syntactic parser is based on the technique given in [21].

We have marked up and verified our own corpus, which consists of 300 "golden standard" sentences collected from classical texts, news and the Internet. Each sentence is unique in its syntax and lexical structure. The length of each sentence ranges in between 2–15 words. We specifically chose sentences for our corpus from various thematic sources such as banks, airlines, religion and literature. Also, we made the corpus cover all the most important language constructions, including coordinated and subordinated clauses.

The results have given the F-score of 82% UAS (unlabeled attach score [14]) with the parsing speed of ~ 2.17 Kbytes of plain text per second.

Figure 12 shows shortened assumption of hybrid tree for a sentence with isolation 2–3 and homogeneous nouns 4–6.

```
<S T="أكلت ، عند صديقي ، تفاحتين و برتقالة">
     <VW="أكلت" GV="V">
          < VW ="عند" GV ="Prep ">
               < VW ="صديقي" GV ="N />
          <Coord>
               <Group GrV="N">
                    <VW="تفاحتين" GV="N"
                    <Sepr Token="و" />
                    <VW="برتقالة" GV ="N"
               </Group>
          </Coord>
     </V>
</S>
```

Fig. 12. A hybrid tree for sentence "أكلت[1] ، عند[2] صديقي[3] ، تفاحتين[4] و[5] برتقالة[6]"
"I_ate[1] with[2] my_friend[3] two_apples[4] and[5] orange[6]"

5. Industrial usage

Described Arabic syntactic parser is currently used as an internal component of Dictum's Opinion Mining system (OMS), an application that is used as a primary component of the social media monitoring service named Kribrum [11].

The workflow looks like the following: OMS receives a review on some topic, passes it to the syntactic parser that analyses it and returns corresponding hybrid trees with semantic information marks provided by rule base system back to the OMS. Then OMS works with hybrid trees to get the summary tonality of the review, collects information about positive/negative aspects of the estimation and finally provides all gathered information to the Kribrum, so it becomes available to users.

6. Discussion

In the paper we shared our experience in building industrial-strength rule-based parser for Arabic. As it was mentioned in the Introduction, the majority of parsers for Arabic use data-driven approach, that is why good performance of our rule-based parser presents new experience in Arabic NLP.

The closest plans are the following:

- Make dictionaries that contain terms taking into account regional specificity and rebuild syntactic structure for expansion the opportunities of the analyzer.
- Add new grammatical characteristics as transitivity of verbs and animateness of nouns to make syntactic analyzing of long and complicated sentences more accurate.

References

1. *Al-Taani A., Msallam M., Wedian S.* (2010), A Top-Down Chart Parser for Analyzing Arabic Sentences, Department of Computer Science, Yarmouk University, Jordan.
2. *Attia M.* (2008), Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation, PhD Thesis, School of Languages, Linguistics and Cultures, the University of Manchester.
3. *Attia M., Pecina P., Tounsi L., Toral A., van Genabith J.* (2011), A Lexical Database for Modern Standard Arabic Interoperable with a Finite State Morphological Transducer, Mahlow, Cerstin, Piotrowski, Michael (Eds.) Systems and Frameworks for Computational Morphology. Second International Workshop, SFCM 2011, Zurich, Switzerland.
4. *Attia M., Pecina P., Tounsi L., Toral A., van Genabith J.* (2011), Lexical Profiling for Arabic. Electronic Lexicography in the 21st Century, Bled, Slovenia.
5. *Attia M., Pecina P., Samih Y., Shaalan K., van Genabith J.* (2012), Improved Spelling Error Detection and Correction for Arabic, COLING, Bumbai, India.
6. *Blinov A. A.* (2009), Territorial'nye varianty arabskogo litaraturnogo jazyka i ih otrazhenie v presse, PhD Thesis, Institute of Oriental Studies of the RAS, Moscow.

7. *Buckwalter T.* (2004), Buckwalter Arabic Morphological Analyzer Version 2.0, Linguistic Data Consortium, Philadelphia, USA.
8. *Eisner J.* (1998), Three new probabilistic models for dependency parsing: An exploration, In: Proceedings of the 16th International Conference on Computational Linguistics (COLING).
9. *Green S., Manning C. D.* (2010), Better Arabic Parsing: Baselines, Evaluations, and Analysis. In COLING 2010.
10. *Habash N. Y.* (2010), Introduction to Arabic Natural Language Processing, Morgan & Claypool, Toronto.
11. Kribrum service website, http://www.kribrum.ru
12. *Kulick S., Gabbard R., Marcus M.* (2006), Parsing the Arabic Treebank: Analysis and Improvements, Treebanks and Linguistic Theories 2006.
13. *Marton Y., Habash N. Y., Rambow O.* (2010), Improving Arabic dependency parsing with lexical and inflectional morphological features, Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages.
14. *McDonald R., Pereira F., Ribarov K., Hajic J.* (2005), Non-projective dependency parsing using spanning tree algorithms, In Proc. of the Joint Conf. on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP).
15. Penn Arabic Treebank project, http://www.ircs.upenn.edu/arabic/
16. *Petrov S., Barrett L., Thibaux R., Klein D.* (2006), Learning Accurate, Compact, and Interpretable Tree Annotation, Proceedings of COLING-ACL 2006.
17. *Shamshad A.* (2012), CYK Algorithm, International Journal of Scientific Research Engineering & Technology (IJSRET), Volume 1 Issue 5.
18. *Smrz O.* (2007), Functional Arabic Morphology: Formal System and Implementation, Doctoral Thesis, Institute Of Formal and Applied Linguistics,
19. Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic.
20. *Skatov D. S., Gergel V. P.* (2013), Efficient Storage Structure Of A Dictionary With String Keys And Associated Values, Vestnik Nizhegorodskogo universiteta im. N. I. Lobachevskogo, Nizhnij Novgorod, Russia.
21. *Skatov D. S., Liverko S. V., Okatiev V. V., Strebkov D. Y.* (2013), Parsing Russian: a Hybrid Approach, Association for Computational Linguistics (ACL), Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing.
22. *Toldova S., Sokolova E., Astaf'eva I., Gareyshina A., Koroleva A., Privoznov D., Sidorova E., Tupikina L., and Lyashevskaya O.* (2012), Ocenka metodov avtomaticheskogo analiza teksta 2011–2012: sintaksicheskie parsery russkogo jazyka [NLP evaluation 2011–2012: Russian syntactic parsers]. In Computational linguistics and intellectual technologies. Proceedings of the International Workshop Dialogue'2012. Vol. 11 (18), Moscow, Russia.
23. *Tounsi L., Attia M., van Genabith J.* (2009), Parsing Arabic Using Treebank-Based LFG Resources, LFG09: 14th International LFG Conference, Trinity College, Cambridge, UK.
24. *Toutanova K., Manning C. D.* (2000), Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger, In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000).

RU-EVAL-2014: EVALUATING ANAPHORA AND COREFERENCE RESOLUTION FOR RUSSIAN

Toldova S. Ju.[1,2] (toldova@yandex.ru),
Roytberg A.[1, 3] (cvi@yandex.ru),
Ladygina A. A.[2] (aladygina@yahoo.com),
Vasilyeva M. D.[2] (linellea@yandex.ru),
Azerkovich I. L.[2] (iazerkovich@gmail.com),
Kurzukov M.[2] (mkurg@ya.ru),
Sim G.[2] (sim.ge@yandex.ru),
Gorshkov D. V.[2] (d.gorshkoff@gmail.com),
Ivanova A.[2] (ivanastas@gmail.com),
Nedoluzhko A.[4] (nedoluzhka@gmail.com),
Grishina Y.[5] (jul_gr@mail.ru)

[1] Natioanl Research University Higher School of Economics, Faculty of Philology, Myasnitskaya 20, 101000 Moscow, Russia

[2] Moscow State University, Philological Faculty, Dept. of Theoretical and Applied Linguistics, Leninskie gory, GSP-1, 119991 Moscow, Russia

[3] IMPB (Institut of mathematical problem in biology) RSS

[4] Charles University in Prague

[5] Applied Computational Linguistics, University of Potsdam

The paper reports on the recent forum RU-EVAL – a new initiative for evaluation of Russian NLP resources, methods and toolkits. The first two events were devoted to morphological and syntactic parsing correspondingly. The third event was devoted to anaphora and coreference resolution. Seven participating IT companies and academic institutions submitted their results for the anaphora resolution task and three of them presented the results of the coreference resolution task as well. The event was organized in order to estimate the state of the art for this NLP task in Russian and to compare various methods and principles implemented for Russian. We discuss the evaluation procedure. The anaphora and coreference tasks are specified in the present work. The phenomena taken into consideration are described. We also give a brief outlook of similar evaluation events whose experience we lay upon. In our work we formulate the training and Gold Standard corpora construction guidelines and present the measures used in evaluation.

Keywords: NLP evaluation, anaphora/coreference resolution, scoring metrics, coreference corpora

1. Introduction

The NLP Evaluation forum RU–EVAL started in 2010 as a new initiative aimed at independent evaluation of NLP systems for Russian. The Third evaluation campaign (2011–2012) focuses on anaphora and coreference resolution. The main objective of the Forum is to promote the development of language technologies for Russian. It unites separate teams dealing with NLP for Russian both from academic institutions and from industrial companies; provides the unified platform for the evaluation of technologies and algorithms; suggest the expertise for the current state-of-the-art in the field.

The organization of the forum is based on the experience of independent NLP systems evaluation events for different languages as well as multilingual evaluation events such as MUC (Message Understanding Conference), EVALITA (evaluation for NLP systems in Italian), ConLL and some others. This campaign is a pilot event for anaphora/coreference resolution tasks for Russian held for the first time. Thus, on the one hand we follow the basic principles of data design and evaluation worked out for afore mentioned events. On the other hand, we used a modified (simplified) conditions.

The forum also has an educational component: the expert group includes students and postgraduates in computational linguistics. It is a good opportunity for them to have a hands-on experience of how the NLP tools work.

The first NLP Evaluation forum focused on morphological taggers (see http://rueval.ru/news.html, [Lyashevskaya et al. 2010], bringing together 15 participants from Moscow, Saint-Petersburg, Yekaterinburg, Ukraine, Belarus and UK. In 2011–2012, syntactic parsing technologies were evaluated [Toldova et al. 2012]. Seven participants took part in the campaign. The results of four participants are available as parallel Treebank now (URL: http://otipl.philol.msu.ru/~soiza/testsynt/).

The present campaign is devoted to two tasks such as coreference chains extraction and anaphora resolution. The main aim of the tasks is to track pronominal or all the mentioning of one and the same entity through the text. The anaphora/coreference resolution modules are important components for the Information extraction systems (named entities recognition and fact extraction tasks in particular) as well as for the MT systems. These NLP components could improve the results for the text summarization and text classification tasks.

The aim of both anaphora and coreference resolution components of an NLP system is to find all the mentions in the text that refer to the same real-world entity, e.g. (1):

(1) *a) Probovali sravnivat' text* ***zapisnoj knizhki***$_x$ *s rukopisjami* ***Nahimova***$_y$ <... >
b) issledovatelej zainteresovala eshcho ***odna jego***$_y$ ***zapisnaja knizhka***$_z$, ***kotoraja***$_z$ *hranitsya sejchas v sevastopol'skom musee. c) V otlichije ot* ***nashej***$_x$ ***eta knizhka***$_y$ (***knizhka***$_y$) *sohranilas' luchshe. d) Na oborote oblozhki* $\varnothing_y$ *rukoj* ***Nahimova***$_y$ *napisano:* ***Pavla Stepanovicha Nahimova***$_y$. … *e) Itak somnenij ne bylo.* ***Obe knizhki*** *prinadlezhali* ***admiraly Pavlu Stepanovichu Nakhimovuy***.
lit. … (they) *tried to compare* (the) *text of* (the) *notebook*$_x$ *with* (the) *Nahimov's manuscripts…The researches were interested in one more* (his) *notebook*$_y$ *kept now in the Sevastopol museum. Unlike ours*$_x$ *this book*$_y$ (book$_y$) *was preserved in better conditions* (as compared to ours). *It is written*

in Nahimov's hand on the reverse cover of this book: …Thus there was no doubt that both of the books had belonged to Admiral Pavel Stepanovich Nahimov.

In (1) we have five Noun Phrases (NP) referring to the same entity Pavel Stepanovich Nakhimov. The first, the third and two others are proper names (the surname Nakhimov, the full name Pavel Stepanovich Nakhimov and the military rank plus full name correspondingly) while the second one is a possessive pronoun. We have also another chain of NPs referring to an entity: these are *eshcho odna jego*$_{y}$ *zapisnaja knizhka, eta knizhka, kotoraja*. Besides full noun phrases there are pronouns such as *jego* 'his', *kotoraja* 'that'. These pronouns have no meaning by themselves, their interpretation dependents on previous expressions in context—its antecedent. Thus we have two types of problems in referent tracking in a text. The first one is the task to gather all the mentions of a referent in a text, using semantic, syntactic and other properties of corresponding NPs. The other one is while seeing an anaphoric element with no semantic clue for its referent to find out what particular noun phrase mentioned in the previous context could serve as such a clue.

We have two corresponding tracks for the present campaign: the coreference resolution task and the anaphora resolution task. The first one presupposes the detection of all the entity mentions and hence gathering all the NPs referring to a particular entity into a chain. The second task was to detect an antecedent of a pronoun in the text and hence to enumerate all two-element chains <antecedent, pronoun>.

This is the first pilot run of the tracks for Russian. Thus, the tasks were limited to the non-event anaphora; no implicit relations between corresponding NPs (such as part-whole, team-member etc.) were involved (see section 3.3 for details).

Both tasks are much more complicated than previous ones (syntactic and morphological annotation). The mainstream technologies in this field presuppose the morphological and syntactic analysis as pre-processing stages. The machine learning techniques are widely used with these NLP tasks.

There were three participants in the first track and seven participants with total 17 runs for the anaphora resolution track. The participating systems vary in their final purposes. Some of them were just experimental systems whose goal was to test some particular anaphora resolution techniques. Other participants are the full-scaled NLP systems having the anaphora/coreference resolution module as its component. For each system has its own NLP pipeline and no generally accepted standards of morphological and syntactic annotation exists for Russian no prerequisite information concerning noun phrase structure or morphological properties was given to the participants. A little manually annotated training corpus consisting of nearly one hundred texts was suggested to the participants in order to give the opportunity to the teams to test various machine learning technique.

The overall procedure was organized as follows: participants received a text collection, processed it in their systems and sent the result back in a unified format. Standard metrics such as precision, recall and F-measure were computed for the anaphora resolution and three types of metrics (section 4) used in reference resolution were assessed by comparing the result against the manually tagged Gold Standard (GS). The expertise of the task output was performed semi-automatically with double manual

check of dubious cases. The set of coreference chains included mostly the NPs referring to the real-life entities (specific referents).

The evaluation procedure has manifested that there are systems for Russian that have quite high precision for the anaphora resolution task though for many systems the recall is low. The coreference task is more complicated. The general problems are as follows.

For languages like Russian the standard methods elaborated for English are not enough due to such features as free word order and no overtly expressed definiteness of a noun phrase.

2. Related evaluation campaigns and datasets

During the organization we relied upon similar evaluation events: MUC-7 [Hirschmann 1997], EVALITA [Uryupina et al. 2011], ARE (Anaphora Resolution Exercise) [Orasan et al.2008], SemEval 2010. First, we need to identify anaphoric and coreferential types which were assessed in these projects. In this section we give brief overview of evaluation initiatives, possible task definitions, available data resources

2.1. MUC

One of the first evaluation events for the tasks under discussion took place within the Message Understanding Conferences (MUCs). The following definition for the coreference relations was suggested: "The coreference 'layer' links together multiple expressions designating a given entity". Only links between noun phrases were considered. The main criteria for the task definition were the good interannotator agreement, the simplicity and speed of text annotation, the objective to create a corpus for independent research of coreference. The MUC project is the best-known example of coreference annotation, on which much subsequent work is based. The various systems results on the test sets from MUC-6 (1995) and MUC-7 (1998) coreference corpora are widely discussed in the literature (c.f. [Mitkov 1999], [Van Deemter, K., & Kibble, R. 2000] and others). The main principle for data annotation was the referential NP identity. We followed the basic MUC data annotation principles.

2.2. Anaphora Resolution Excersise 2007 (ARE)

The other influential event in the coreference resolution domain is Anaphora Resolution Exercise[1] that was held within the Discourse Anaphora Colloquium. There were four tracks depending on the anaphora types and the types of prior information. The data sets for the first two tracks were pre-annotated: the NPs were detected and tagged, some of them were tagged as NPs for which the detection of the antecedent is required. The data for two other tracks remained unannotated.

[1] Discourse Anaphora and Anaphor Resolution Colloquium (DAARC2007)

2.3. SemEval 2010

The SemEval-2010 task is of special interest for us in the following respect: it is the task task on Coreference Resolution in Multiple Languages (six languages such as Catalan, Dutch, English, German, Italian and Spanish). Besides the well-studied languages, the under-resourced languages took part in the anaphora resolution procedure. Some of the participants used small training and testing data sets (e.g. 80 texts for Italian as a training set and 46 texts as a test set). The other issue of interest is that four different metrics were used in evaluation. The corpora were annotated with morphological, syntactic and partial semantic tags. One of the aims of the event was to learn out what was the impact of different levels of linguistic information.

2.4. Evalita 2011

In the latest anaphora resolution tasks such as Evalita-11 task the systems are required to recognize all the mentions of an entity and to cluster them into a chain irrespective of the NP type (including names, pronouns, zero pronouns, etc.) The singleton NPs are also included. Not only referring NPs are taken into consideration. All the NPs were marked and annotated [Uryupina et al. 2011]. The additional information on morphological and syntactic token properties was given as well as some mention and semantic types.

2.5. Some important corpora

Our data set includes training corpus and Gold Standard corpus that were designed considering the existing linguistic corpora for other European languages which contain coreferential annotations: OntoNotes [Hovy et al., 2006; Pradhan et al., 2007], Potsdam Commentary Corpus [Stede, 2004], ARRAU Corpus [Poesio & Artstein, 2008], Prague Dependency Treebank [Böhmová et al., 2003], VENEX Corpus [Poesio et al., 2004]. In our work we tried to rethink the existing methodology and implement it to Russian.

We also studied the annotation principles and annotation schemes used for coreference corpora presented in [Nedoluzhko et. Al 2009], [Nedoluzhko 2013], [Khudyakova 20011]. However these corpora are based on deep annotation schemes. The first one requires the syntactic analysis as the prerequisite condition. It also deals with implicit anaphoric relations such as bridging. The detailed multifactor annotation scheme is implemented for the latter corpus.

Thus our annotation scheme was based on MUC corpora principles. For the simplification of annotation task only identity relations were annotated, other types like bridging anaphora, predicative anaphora, coreference of events were excluded. We also decided to ignore these types of coreferential relations based on the results of several experiments to evaluate the capacity of existing Russian coreference resolvers.

Unlike SevEval 2010, we do not take into account expressions that occurred only once. We did not take into account morphological and syntactical tags (despite the fact that we included them into the output to make the manual evaluation easier).

3. Participants and data sets

3.1. Tasks

As it was mentioned above the tasks proposed in this exercise focused on problems related to anaphora and coreference resolution.

The purpose of the first task was to evaluate the quality of different pronoun resolution algorithms. The systems were expected to recognize pronouns which need to be resolved (we examined only personal, possessive, reflexive pronouns and a relative pronoun *kotoryj* 'which' as well) and find their antecedents. We use the lenient principle of evaluation assuming the transitivity of anaphoric relations. The antecedent NP established by a system should be a member of corresponding coreference chain in the Gold Standard set.

In the second task the participants were required to recognize referential expressions (proper names, nouns and pronouns, excluding zeroes) in an unannotated document and cluster them into coreference chains.

3.2. Participants

Eight NLP groups from Moscow and St. Petersburg expressed their interest in participating in the tracks of the Third RU-EVAL forum. They are: Compreno (Abby, Moscow), RCO (Russian Context Optimizer, Moscow), SemSyn K. Boyarsky, E. Kanevsky, St. Petersburg). Open Corpora (St.Petersburg), Mail.ru anaphora resolver (Mail.ru, Moscow), the system of Institute for Systems Analysis of Russian Academy of Sciences (ISA RAS), Sergej Ponomarev's system. Three teams took part in coreference resolution task as well. Some teams submitted results for different machine learning and rule based techniques. Thus, we have 17 answers from 7 systems for anaphora resolution track.

3.3. Main principles of dataset preparation

While preparing the data sets for the tracks we based on the principles of the MUC anaphora resolution corpora creation. The resulting corpus is the first open corpus annotated for coreference relations for Russian, and thus, the main purposes and principles of our corpus constructing are: the resulting corpora should be of open access in order to be freely distributed among the community;

1) the resulting corpora should be open access in order to be freely distributed among the community;

2) it should be reusable, open access, distributed in a machine readable format on the one hand and presented in human readable form on the other hand (the platform for easy annotation and visualization of data should be provided);
3) the corpus should include various genres (in contrast to the majority of coreference corpora);
4) the annotation procedure should be simple,
5) the high annotator agreement should be the crucial criteria for chain inclusion into the corpus;
6) only entity referring expressions are considered (no event anaphora);
7) only identity relation between NPs referents (expanded to near-identity in some cases) are under consideration (no bridging, split referents etc.).

3.4. Training and test corpora

This project required work with the widest possible selection of texts. For this reason, short texts or fragments of texts in a variety of genres have been included in the test corpus: news, scientific articles, blog posts and fiction. All texts were taken from publically available sources, such as Russian OpenCorpus, online library Lib.ru and Lenta.ru. For test corpus about 300 texts were used from each above mentioned source, to the total of 1342 texts. 20 texts of each section were used for annotating the Golden Standard.

In preparation for the Anaphora Resolution Event, a subcorpus of approximately 100 texts was selected as a training corpus for the participants. The texts were tokenized, split into sentences, pos-tagged with TreeTagger for Russian (we used a TreeTagger-based ([Schmid 1994]) part-of-speech tagger, a lemmatizer based on CSTLemma ([Jongejan, Dalianis 2009]) available at URL: http://corpus.leeds.ac.uk/mocky/).

These texts were used as a basis for training the participating systems, which later were tested against the whole corpus. The texts were from 5 up to 100 sentences long, the longest one being 170 sentences long. In total 2000 anaphoric pronoun—antecedent pairs and 1200 coreferential chains were annotated by hand. The texts were annotated by 2 annotators and the differences between their annotations were compared and analyzed. After that they were checked by supervisors who made the necessary updates.

The special Web-interface was designed by Dmitrij Gorshkov for corpus annotation. The tool uses MySQL database engine for corpus management. While constructing this tool we took into account some features of MMAX-2 mark-up tool (available at http://mmax2.sourceforge.net/, see [Müller C., Strube M. 2003]), Brat annotation tool (available at http://brat.nlplab.org/, see also [Nilsson Björkenstam, K., & Byström, E. 2012]) and some others. However we decided to use our own tool based on MySQL since it provided the flexible work with corpus data, convenient visualized Web-interface suitable for collaborative work on corpus annotation and the annotation comparison.

3.5. Gold Standard Preparation: basic principles and instructions

In the corpora (training and Gold Standard), the following NPs were annotated:
(1) pronominal and nominal NPs referring to real-world entities;
(2) non-specific (generic and abstract NPs) if they are antecedents of pronouns from our list (tagged for the anaphora resolution track only)

According to the results of several experiments on assessment of the capacity of existing Russian coreference resolvers, we decided to ignore:
1) bridging relations (c.f. part-whole relation in *zapisnaja knizhka* 'the note-book' and *oblozhka* 'the cover (of the note-book)' in example (1)),
2) discourse deixis (1st and 2nd person pronouns)
3) and coreference relations with a split antecedent in our annotation;
4) discontinuous expressions (c.f. "***Peter*** *came there with* ***Masha****. They…*") and split antecedents.

Development of our mark-up scheme was influenced by the annotation guidelines proposed in [Krasavina and Chiarcos, 2007]. These guidelines were created for annotating coreference in German and English. We used slightly modified annotation scheme and our own Guidelines (adapted for situation with Russian NLP systems).

Referential expressions subject to annotation are called markables (or groups in our corpus). Only NPs can be markables. Markables are maximal NPs excluding the relative clauses, the postpositional participle constructions and some other (see above). The appositive NPs are treated as separate markables (groups).

There is also a distinction between primary markables and secondary markables. Thus, following expressions should be annotated as primary markables:

1) 3rd person pronouns (1st person pronoun is marked if only it denotes the narrator in the text);
2) demonstrative pronouns;
3) reflexive pronouns;
4) possessive pronouns;
5) definite and possessive descriptions;
6) proper names and titles;
   Some other markables, which are not annotated as primary markables, but are potential antecedents for them, are also considered to be secondary markables, e.g.:
7) indefinite descriptions if they are used for the first mentioning of an entity.
8) apposition NPs as *Admiral* in

(2) *Pavel Stepanovich Nakhimov, vydajuscshijsya* ***rossijskij admiral—***
'Pavel Stepanovich Nakhimov, 'the great Russian Admiral';

9) predicative NPs as *Gallej* in

The comet was called ***Gallej*** *or as in He became a* ***great scientist****;*

10) first and second person pronouns in the direct speech constructions.

The main difference between primary and secondary markables is that the former are always annotated while the latter are annotated only if they potentially could serve as antecedents for some of primary markables. The secondary markables are do not participate in the evaluation score. However the system is not penalized for establishing coreference relations with them.

For each markable there is a number of attributes to be defined. These are, for example, functional type: 'def'—for the expressions referring to the entities, 'pred'—for the predicative referring expressions, 'appo'—for appositive expressions, 'ds'—for 1st and 2nd person pronouns in direct speech, 'misc' for some other dubious cases such as near-identity cases (c.f. *his book—this edition of the book*). We also use the tag 'meton' for metonymies for it helps to single out the cases that are highly difficult for the detection of coreference. We also use special set of attributes for NP structure: the separate attributes for different types of pronouns ('refl', 'dem', 'rel' etc.) and 'noun' for non-pronominal NPs.

The process of annotation is based on several principles, also described in the guidelines. These principles refer to the order of establishing coreferential links and forming coreferential chains as well as annotating markables of maximal size. For each group longer than one word we mark the potential semantic head. There were participants who detected only heads as referring expressions. Moreover the NP heads could vary through systems. Thus, in the example in (2) the head could be Pavel, Nahimov, admiral. The information about potential heads is used in the evaluation procedure (see below).

The data for training include: 1) the group (markable) offset presented as the shift from the beginning of a given text in symbols and the length of the fragment; 2) its ID, 3) the text ID; 4) the chain ID 5) the referential expression itself.

The training data was distributed as a set of plain texts and an xml-file with anaphoric chains information:

1) in the anaphora dataset a chain consists of two elements: a pronoun from a list of pronouns (3rd person, possessive 3rd person and reflexive, demonstratives and the relative pronoun *kotoryj* 'that');
2) in the coreference dataset a chain consists of all the NPs—mentions of the same entity with a set of attributes in the training corpus and without attributes in the testing set.

The same format is used for the systems response set.

This information was used both for automated comparison with the Golden Standard and for manual check by the annotators.

4. Measures

Systems were examined in two tasks: anaphora resolution and coreference resolution. We use F-measure to combine recall and precision to evaluate pronoun

resolution algorithms work. We use MUC, B^3 and CEAF to compute a more complicated recall and precision for coreference and then we use F-measure. The results of evaluation are published at http://rueval.compling.net/anaph/results.html.

Below we use GOLD for Gold Standard corpus chains (linked groups) and RESPONSE for the system response chains and groups. We use the principle of lenient groups matching (the same for both tracks): the boundaries of GOLD NP should intersect the boundaries of System NP. The GOLD NP head should be included into intersection. A System NP matches the only one GOLD NP. There is an exclusion from the rule: when System NP includes the GOLD NP and the head noun of the System NP is out of the GOLD NP there is no groups match. The inaccurate NP boundaries (as in *oblomok skaly v reku*—'the piece of the rock (fell)') are not penalized.

4.1. Measures for anaphora resolution (precision, recall, f-measure)

We used standard measures for anaphora resolution track. These are precision, recall and F-measure.

We use the "lenient" principle for assessment of matching the system response to the Gold Standard anaphoric pair. The true set of entities is a set of all pronouns enumerated in 3.5. in GOLD with their antecedents. We map the System chains to GOLD chains according to the following principle: a chain from GOLD that includes a particular pronoun corresponds to the chain in RESPONSE with the same pronoun.

We consider the system response True Positive if the chain in RESPONSE with a particular pronoun is a subset of GOLD chain with the same pronoun (lenient group matching principle).

Precision then shows what proportion of System pairs match the GOLD (what part of all true pronouns and true links was found). S is all pronouns with their link from system output and M is all true pronouns with true links from system output. So we can compute the precision:

(1) $P=M/S$.

Recall shows what part of system output is true pronouns with true links, to compute this we use the formula:

(2) $R=M/G$.

The formula for F-score is:

(3) $F=2PR/P+R$,

where F is F-score, R is recall and P is precision.

4.2. Measures for coreference resolution MUC, B-Cubed, CEAF

We use three measures for the coreference track evaluation: MUC, B^3 and CEAF.

4.2.1. MUC-score

MUC is a link-based metric (for MUC measures see e.g. [Chen & Ng, 2013], [Vilain et al.1995] etc.). Recall and precision are associated with links between the golden standard chains and the system chains.

Recall is computed as the number of common links between the golden chains and the system chains in a document divided by the number of links in the golden chains.

Precision is computed as the number of common links divided by the number of links in the system chains.

4.2.2. B-cubed

B-cubed is entity-oriented cross-document coreferencing measure ([Bagga and Baldwin, 1998]). B-cube is a more complicated harmonic mean of recall and precision.

To compute B^3 one needs to compute the $recall_i$ and $precision_i$ for each mention, and then to take an average of these pre-recall and pre-precision values to obtain the overall recall and precision.

Mention precision is a number of correct elements in the system output chain containing the mention divided by the number of elements in the system output chain containing the $mention_i$

Mention $recall_i$ is a number of correct elements in the system output chain containing the mention divided by the number of elements in the golden standard chain containing the $mention_i$.

Hence, to compute overall recall and overall precision you need to average all mention recalls and mention precision respectively.

4.2.3. CEAF

In [Luo 2005] it was shown the main problem of B^3 algorithm: B^3 may use all chains more than once when computing recall and precision. Luo proposes 2 metrics which aligns entities in golden standard and system output. First of all CEAF requires the establishing of all one-to-one corresponding alignments between the chains in G(d) and the chains in S(d). CEAF uses the function *j* to compute the similarity between G_i and S_j where G_i and S_j are the chains from golden standard and system output respectively. Furthermore, the algorithm proposes the best alignment using the Kuhn-Munkres algorithm.

4.2.4. The Evaluation procedure

Taking the above described measures into account we made the revision of Gold Standard data. First of all we checked the consistency of annotators principles and "switched off" all the chains that seemed not referential or caused some problems and discussions between annotators. Then we implemented the procedure of groups mapping and chain mapping. A random sample of mapped groups were checked manually. A random sample of erroneous links as well as missed links was also manually checked. Then the corresponding scores were automatically calculated.

Conclusions

The RU-EVAL 2014 has brought together a number of IT companies and academic groups that work on Russian anaphora and coreference resolution, and made it possible to assess the state-of-the-art in the field (so far, mostly in Russia). The forum has shown that there are competitive teams that develop high-level (discourse level) NLP components on a considerably high level (some systems manifest nearly 80% precision for anaphora resolution). However, the task of anaphora resolution is complicated for Russian due to free word order and the absence of overt markers of NP referential status. The absence of free semantic resource as WordNet and freely distributed syntactic parsers make the task more difficult for NLP start-ups and new small teams. The anaphora and coreference resolution tracks have shown the impact of high quality lower level linguistic analysis to the quality of discourse analysis tasks. However the event was the challenge for those teams that conduct the experiments on various machine learning techniques.

The event has the following practical outcomes:

- the baseline for anaphora and coreference resolution for Russian was evaluated
- the guidelines for tagging according to GS principles have been compiled and tested for Russian;
- new anaphora resolution systems for Russian arises at stretch due to the RU-EVAL 2014 campaign;
- the manually tagged standard set, consisting of nearly 200 texts annotated for anaphora and coreference chains is made available through http://gs-ant.compling.net/ and http://ant.compling.net/ (the latter is to be moved to the former URL);
- the created corpus includes a wide variety of genres and various types of coreference relations.

The organizers hope that these corpora would be helpful for other NLP teams for the experiments on coreference resolution algorithms.

Acknowledgments

We would like to thank other MSU students who participate in corpora preparation: Darya Skorobogatova, Alexander Pechonyj, Viktoria Danilova, Alexander Kostyuk, Max Ionov. We also are grateful to them and to some evaluation event participants: Anna Sergeeva, Natalya Glazyrina, Ivan Hramoin for their help in training corpus annotation. We also would like to thank Olga Lyashevskaya, Anastasija Bonch-Osmolovskaya for their help in the event organization. We are also most grateful to the participants of the forum.

The work was partially funded by the LINDAT/CLARIN project (project LM2010013).

References

1. *Bagga, A., Baldwin B.* (1998). Algorithms for scoring coreference chains, Proceedings of the 1st International Conference on Language Resources and Evaluation, Granada, Spain, 28–30 May 1998, pp. 563–566

2. *Böhmová, A., Hajič, J., Hajičová, E., & Hladká, B.* (2003). The Prague dependency treebank. In Treebanks, pp. 103–127. Springer Netherlands.
3. *Chen Ch., Ng V.* (2013). Linguistically Aware Coreference Evaluation Metrics, Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP-13), 2013, available at: http://www.hlt.utdallas.edu/~vince/papers/ijcnlp13-coref.pdf.
4. *Dipper, S., Zinsmeister H.* (2012), Annotating Abstract Anaphora. Language Resources and Evaluation 46 (1), pp. 37–52.
5. *Gareyshina Anastasia, Ionov Maxim, Lyashevskaya, Olga, Privoznov Dmitry, Sokolova Elena, Toldova Svetlana.* (2012). RU-EVAL-2012: Evaluating Dependency Parsers for Russian. Proceedings of COLING 2012: Posters. pp. 349–360. URL: http://www.aclweb.org/anthology/C12–2035.
6. *Hirschmann L.* (1997). MUC-7 coreference task definition. Version 3.0, Proceedings of the 7th Message Understanding Conference (1997).
7. *Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R.* (2006, June). OntoNotes: the 90% solution. In Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers, pp. 57–60. Association for Computational Linguistics.
8. *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. Suntec, Singapore : Association for Computational Linguistics, 2009. pp. 145–153.
9. *Khudyakova, M. V., Dobrov, G. B., Kibrik, A. A., & Loukachevitch, N. V.* (2011). Computational modeling of referential choice: Major and minor referential options, Proceedings of the CogSci 2011 Workshop on the Production of Referring Expressions. Boston (July 2011).
10. *Krasavina, O., Chiarcos Ch.* (2007) PoCoS: Potsdam coreference scheme. Proceedings of the Linguistic Annotation Workshop. Association for Computational Linguistics, 2007.
11. *Luo, X.* (2005, October). On coreference resolution performance metrics. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (pp. 25–32). Association for Computational Linguistics. Available at: http://dl.acm.org/citation.cfm?id=1220579.
12. *Lyashevskaja Olga, Astaf'eva Irina, Bonch-Osmolovskaja, Anastasia, Gareyshina Anastasia, Grishina Julia, D'jachkov Vadim, Ionov Maxim, Koroleva Anna, Kudrinskij Maxim, Lityagina Anna, Luchina Elena, Sidorova Evgenia, Toldova Svetlana, Savchuk Svetlana., Koval' Sergej.* (2010). Evaluation of the automated text analysis: POS-tagging for Russian. [Morphological Ananlysis Ocenka metodov avtomaticheskogo analiza teksta: morfologicheskije parsery russkogo jazyka.] Proceedings of the International Conference on Computational Linguistics Dialogue-2010. [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2010"], pp. 318–327.
13. *Mitkov, R.* (1999). Anaphora resolution: the state of the art. School of Languages and European Studies, University of Wolverhampton., available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.29.6235&rep=rep1&type=pdf

14. *Müller C., Strube M.* (2003). Multi-level annotation in MMAX. In Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue.
15. *Nedoluzhko, A.* 2013 How Dependency Trees and Tectogrammatics Help Annotating coreference in Prague Dependency treebank. DepLing 2013, Prague. No. 44, ÚFAL, Charles University in Prague.
16. *Nedoluzhko, A., Mírovský, J., & Pajas, P.* (2009, August). The coding scheme for annotating extended nominal coreference and bridging anaphora in the Prague Dependency Treebank, Proceedings of the Third Linguistic Annotation Workshop, pp. 108–111. Association for Computational Linguistics.
17. *Nilsson Björkenstam, K., & Byström, E.* (2012). SUC-CORE: SUC 2.0 Annotated with NP Coreference. In Proceedings of the Fourth Swedish Language Technology Conference (SLTC), October 24–26, 2012, Lund.
18. *Orasan C., Cristea D., Mitkov R., and Branco A.* (2008) Anaphora resolution excersice: an overview, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, May. 2008.
19. *Poesio, M., and Artstein R.* (2008) Anaphoric annotation in the ARRAU corpus, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, May. 2008.
20. *Poesio, M., Delmonte, R., Bristot, A., Chiran, L., & Tonelli, S.* (2004). The VENEX corpus of anaphora and deixis in spoken and written Italian. University of Essex.
21. *Schmid, H.* (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees, Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.
22. *Stede, Manfred* (2004) The Potsdam commentary corpus, Proceedings of the 2004 ACL Workshop on Discourse Annotation. Association for Computational Linguistics, 2004.
23. *Uryupina O., Poesio M.* (2011). Evalita 2011. Anaphora resolution task. In Proceedings of Evalita 2011.
24. *Van Deemter, K., & Kibble, R.* (2000). On coreferring: Coreference in MUC and related annotation schemes, Computational linguistics, 26(4), pp. 629–637.
25. *Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly & Lynette Hirschman* (1995). A model-theoretic coreference scoring scheme. In Proceeding of the 6th Message Understanding Conference (MUC-6), pp. 45–52. San Mateo, Cal.: Morgan Kaufmann.

О ПРОИЗВОДНЫХ ПРЕДЛОГАХ: НАРЕЧНЫЕ ПРЕДЛОГИ[1]

Урысон Е. В. (uryson@gmail.com)

Институт русского языка им. В. В. Виноградова
РАН, Москва, Россия

Ключевые слова: наречие, предлог, невыразимый семантический актант, синтаксический актант, синтаксическая сфера действия, теория валентностей

ON DERIVED PREPOSITIONS: ADVERBAL PREPOSITIONS

Uryson E. V. (uryson@gmail.com)

Vinogradov Russian Language Institute
of the Russian Academy of Sciences, Moscow, Russia

The object of this paper is so called adverbal prepositions in Russian; such as VOKRUG (kostra) ‘around smth.’, DALEKO OT (doma) ‘far from smth.’, etc. By definition, an adverbial preposition coincides with an adverb (cf. VOKRUG) or contains an adverb and a preposition (cf. DALEKO OT). In most cases, an adverbial preposition and the underlying adverb have the same meaning and the same semantic actants. The only difference between an adverbial preposition and the underlying adverb is the mode of expression of the main semantic actant. Cf. GOREL KOSTER, VOKRUG (preposition) KOSTRA STOJALI LIUDI ‘A fire was burning, people were standing around it’ vs GOREL KOSTER, VOKRUG (adverb) STOJALI LIUDI ‘A fire was burning, people were standing around’. Both the adverbial preposition VOKRUG and the adverb VOKRUG have a semantic actant ‘reference point’ and in both examples the word ‘fire’ expresses this actant. But the adverbial preposition governs this noun predicting its case-form and its linear position in a sentence. The adverb does not govern the noun; the only requirement is that this object must be already mentioned (so the noun must be somewhere in the preposition to the adverb). In this regard the adverbs under discussion are similar to connectors. Adverbial prepositions are easily described in the frameworks of valency theory. I argue that some refinements of valency theory are necessary for representing syntactic properties of underlying adverbs. I also demonstrate that it is more convenient to represent so called adverbal prepositions as adverbs but not as prepositions.

Keywords: adverb, preposition, semantic actants, syntactic actants, valency thery

[1] Работа поддержана грантами ОИФН, РГНФ 10-04-00273а и НШ-6577.2012.6.

0. Введение

Русская грамматика традиционно делит предлоги на первообразные и непервообразные, или производные. Примеры первообразных предлогов: *на, в, о, из* и т. п. В эту группу входят также «парные предлоги — сращения» *из-за, из-под* и т. п. [Русская грамматика 1980: 707]. Первообразные предлоги представляют собой закрытую немногочисленную группу «простейших слов» [Русская грамматика 1980: 706]. Эта группа закрыта в том смысле, что она не пополняется. Иными словами, первообразные предлоги могут быть заданы списком.

К производным предлогам традиционная грамматика относит формы некоторых полнозначных слов, а также целые сочетания слов. Примеры: *вокруг (дома), далеко от (дома), благодаря (яркому солнцу), несмотря на (дождь), в связи с (реконструкцией)*. Группа производных предлогов очень многочисленна и разнородна. Она открыта, т. е. в языке образуются новые производные предлоги: есть единицы, «приобретающие свойства предлогов» [Русская грамматика 1980].

Критерии, которые позволяли бы отличить непервообразный предлог от слова или сочетания слов, в грамматике в явном виде не формулируются, т. е. производные предлоги выделяются по интуиции.

По-видимому, главное основание, на котором слово или группа слов зачисляется в производные предлоги, — это его синтаксические свойства: производный предлог, так же как и первообразный, должен «соединять два слова» в словосочетании. При этом управляемая форма всегда находится в постпозиции к предлогу, будь он первообразный или производный. Ср. синтаксическое сходство сочетаний внутри пар: (а) *сидеть У костра — сидеть ВОКРУГ костра;* (б) *дочерна загореть ИЗ-ЗА яркого солнца — дочерна загореть БЛАГОДАРЯ яркому солнцу*. Для определенности будем вслед за МСТ считать, что предлог управляет падежной формой, а сам зависит от другого слова (если оно есть): *сидеть → у → костра*.

Есть и другие свойства, которые сближают производный предлог с обычным предлогом, однако они не так важны для настоящей работы, и мы на них сейчас не останавливаемся.

В идеале требуется явно сформулировать те критерии, по которым слово или словосочетание зачисляется в производные предлоги. Кроме того, нужно понять, насколько принятые, пусть пока имплицитные критерии совместимы с грамматическим описанием других частей речи: как будет ясно из дальнейшего, здесь возможны противоречия, требуется их выявить и понять, какие альтернативные описания возможны в том или ином случае и какое описание предпочтительнее.

В полном объеме поставленная задача пока не выполнена. Объект данной работы — одна группа русских производных предлогов, а именно предлоги, образованные от наречий (наречные предлоги). Мы попытаемся ответить на сформулированные вопросы применительно к этой группе слов.

1.0. Наречные предлоги

По определению, в наречном предлоге обязательно есть компонент, совпадающий с наречием [Русская грамматика 1980]. Примеры: наречный предлог *вокруг кого/чего-л.* имеет в своем составе компонент *вокруг*, совпадающий с наречием *вокруг,* ср. *Вокруг стояли люди*; предлог *вдали от кого/чего-л.* имеет в своем составе компонент *вдали*, который совпадает с наречием *вдали*, ср. *Вдали виднелись горы*.

Ясно, что наречный предлог по своим основным синтаксическим функциям похож на обычный предлог. Ср. *стоять → вокруг → костра; сидеть → далеко от → сцены*.

При этом по данным словарей, некоторые наречные предлоги по семантике совпадают с соответствующими наречиями. Ограничимся несколькими очевидными примерами: *вокруг кого/чего-л.* и *вокруг*; *навстречу кому/чему-л.* и *навстречу*; *напротив кого/чего-н.* и *напротив*; *вдали от кого/чего-л.* и *вдали; вблизи от кого/чего-л. и вблизи*; *рядом с кем/чем-н.* и *рядом; следом за кем/чем-н.* и *следом*.

Такое совпадение значений, вообще говоря, необязательно. Мы, однако, будем рассматривать случай совпадения семантики наречного предлога и производящего наречия — перед нами оказывается своего рода «минимальная пара», помогающая лучше понять логику традиционного грамматического описания.

Если исходить из синтаксических критериев, то наречие и наречный предлог — это две разные лексические единицы. Поэтому толковые словари представляют их как две разные лексемы одного слова[2]. Если же исходить из семантических критериев, то остается неясным, почему наречный предлог, не отличающийся по значению от мотивирующего его наречия, выделяется в отдельную лексему. Действительно, в [НОСС] такое наречие и производный от него предлог рассматриваются в рамках одного значения слова (наречия), см. статьи [Богуславская 2004а; 2004б; 2004в]. Выясним, насколько последовательно каждое описание.

Вслед за [Русская грамматика 1980] нам будет удобно разделить наречные предлоги на простые и составные. «Простые предлоги совпадают с наречием» [Русская грамматика 1980: 707]; ср. *вокруг кого/чего-л.* и т. п. «Составные предлоги представляют собой соединение наречия с первообразным предлогом» [Там же]; ср. *вдали от кого/чего-л.* и т. п. Начнем с простых наречных предлогов.

1.1. Синтаксис наречий и наречных предлогов

Сравним примеры:

(1) *Горел костер, вокруг стояли люди.*
(2) *Вокруг костра стояли люди.*

---

[2] В соответствии со словоупотреблением, принятым в Московской семантической школе, мы называем лексемой слово, взятое в одном из его значений.

В обоих случаях слово *вокруг* предполагает объект X, со всех сторон окруженный другими объектами, а они в свою очередь являются участниками ситуации P. Будем называть объект X ориентиром, а P — ориентируемой ситуацией (участников P можно называть ориентируемыми объектами). Итак, слово *вокруг* в обоих случаях имеет два семантических актанта — ориентир и ориентируемую ситуацию: *вокруг X-а P*. В приведенных примерах оба семантических актанта выражены: объект-ориентир — это *костер*, а ориентируемая ситуация — *стояли люди*.

Различаются примеры (1) и (2) тем, как в них выражен семантический актант ‘ориентир’. В (2) этот актант обозначен формой родительного падежа существительного *костер*: слово *вокруг* управляет формой *костра*, ср. *вокруг → костра*. При этом слово *костер* (точнее, его синтаксическая группа) следует непосредственно за словом *вокруг*. В данном случае выражение семантического актанта ‘ориентир’ слова *вокруг* описывается вполне четким правилом — так задаются синтаксические актанты глагола или предлога. Следовательно, семантическому актанту ‘ориентир’ слова *вокруг* соответствует синтаксический актант. В этом отношении слово *вокруг* в (2) не отличается от обычных предлогов, ср.

(3) *У костра стояли люди.*

По аналогии с обычными предлогами слово *вокруг* в случаях типа (2) тоже считают предлогом.

Иначе обстоит дело с примером (1). Семантический актант ‘ориентир’ слова *вокруг* тоже выражен здесь словом *костер*, но обозначение этого актанта не может быть полностью описано четкими синтаксическими или иными лингвистическими правилами. Относительно выражения этого актанта известно лишь, что он находится где-то в предтексте. Но для его обнаружения недостаточно синтаксических, морфологических или иных жестких правил — требуется привлекать информацию, относящуюся к области понимания текста.

Аналогичным образом обстоит дело и с составными наречными предлогами, ср.

(4) *Во дворе росла ель, рядом стояли две березы.*
(5) *Рядом с елью стояли две березы.*

В обоих случаях слово *рядом* имеет семантический актант ‘ориентир’. Однако в (5) он выражен предложно-падежной группой *с кем/чем-л.*, следующей непосредственно за словом *рядом*, а в (4) выражение того же семантического актанта должно находиться где-то в предтексте, но на его падежную форму не налагаются никакие ограничения. Следовательно, в случае (5) выражение семантического актанта ‘ориентир’ слова *рядом* полностью описывается синтаксическими правилами, подобно тому как задается синтаксический актант обычного предлога. По аналогии с предлогами слово *рядом* в (5) тоже можно было бы считать предлогом. Однако *рядом* в (5) управляет предлогом *с*, а предлог может управлять только существительным. Из-за этого в грамматике предлогом считают все сочетание *рядом с*. В (4) *рядом* — это безусловно наречие; на выражение его семантического актанта ‘ориентир’ накладывается лишь

одно синтаксическое ограничение — обозначение данного актанта должно находиться в предтексте. Следовательно, выражение данного актанта может быть описано синтаксическими или иными подобными правилами лишь частично.

Подобная проблема возникает при описании выражения семантических актантов некоторых союзов и частиц (Урысон 2012; 2013). Приложим приведенное в этих работах рассуждение к нашим случаям.

Семантическому актанту 'ориентир' наречного предлога *вокруг* или *рядом с* соответствует синтаксический актант, причем этот синтаксический актант является обязательным. Обоснуем, что тому же семантическому актанту наречий *вокруг* и *рядом* не соответствует никакой синтаксический актант в обычном понимании этого термина.

По определению, синтаксический актант некоторой предикатной лексемы — это ее синтаксическое зависимое, а оно располагается в пределах того же высказывания, что и сам данный предикат. Обратим внимание на то, что теория валентностей (в частности, используемая в модели «СМЫСЛ↔ТЕКСТ») предназначена прежде всего для представления семантических и синтаксических актантов предиката внутри предложения [Мельчук 1974; Апресян 1974]. Рамками высказывания ограничивается и более широкая теория сфер действия лексических единиц [Богуславский 1996]. Между тем наречия *вокруг, рядом* и т. п. не налагают столь жестких требований на выражение обсуждаемого семантического актанта.

Тем не менее, какое-то формальное требование к его обозначению предъявляется: данный актант может быть выражен только в предтексте. Однако такое требование недостаточно для автоматического анализа или синтеза текста.

В рамках модели «СМЫСЛ↔ТЕКСТ» (да и любой собственно синтаксической концепции) описание подобных ограничений на выражение семантического актанта не предусмотрено. Попытаемся применить к нашему случаю концепцию сфер действия лексических единиц [Богуславский 1996].

2.0. Обсуждение теоретического аппарата: понятие «невыразимого» семантического актанта

Идея И. М. Богуславского состоит в том, что «разумно было бы называть синтаксическим актантом любое синтаксическое образование, значение которого соответствует семантическому актанту» [Богуславский 1996]. Но поскольку термин «синтаксический актант» имеет вполне устоявшееся, традиционное понимание, не распространяющееся на любое синтаксическое образование, то вместо него вводится новый, более широкий термин: синтаксическая сфера действия лексемы по ее семантическому актанту.

Казалось бы, приведенное определение идеально подходит для нашего случая: у наречий типа *вокруг* и *рядом* есть синтаксическая сфера действия по семантическому актанту 'ориентир', а значит, и соответствующий синтаксический актант.

Однако во всех случаях, разобранных И. М. Богуславским, речь идет о тех синтаксических сферах действия, которые задаются четкими, вполне алгоритмизуемыми лингвистическими правилами. Этот факт имеет принципиальный характер: если, заменяя понятие синтаксического актанта более широким и гибким

понятием синтаксической сферы действия, мы откажемся от четких требований к поверхностному выражению актанта, то нам придется отказаться и от идеи алгоритмизуемого обнаружения синтаксического представления высказывания.

Ясно, что выражение семантического актанта ‘ориентир’ наречий типа *вокруг* или *рядом* не может быть задано строгими синтаксическими правилами.

В рамках модели «СМЫСЛ ↔ ТЕКСТ» в таких случаях принято говорить, что данный семантический актант невыразим и лексема не имеет соответствующей синтаксической валентности. Ср. классический пример — глагол *промахнуться* [Мельчук 1974: 135; Апресян 1974: 148].

Глагол *промахнуться* описывает ситуацию, когда субъект стреляет или бросает нечто в цель, или мишень, однако не попадает в нее. Следовательно, у глагола *промахнуться* есть семантический актант ‘мишень’. Однако данный актант не может оформляться как обычный синтаксический актант глагола: в русском языке нет стандартного словосочетания ??*промахнуться по кому-л./чему-л. <*в кого-л./что-л.>*, хотя нормально словосочетание *стрелять в кого-л./что-л.* При этом данный «невыразимый» семантический актант свободно выражается в тексте. Ср. *Охотник выстрелил в волка, но промахнулся*. Однако его выражение невозможно задать теми формальными правилами, которыми описывается реализация синтаксических валентностей[3].

Семантический актант ‘ориентир’ наречий типа *вокруг* и *рядом* тоже не может быть полностью задан формальными синтаксическими правилами. Следовательно, этот актант тоже является «невыразимым».

Однако между глаголом *промахнуться* и обсуждаемыми наречиями есть принципиальное различие.

Семантический актант ‘мишень’ глагола *промахнуться* может вообще никак не называться в тексте, что не нарушает его правильности. Ср. *Не знаю, куда он метил, знаю только, что промахнулся*.

Заметим, что подобным образом устроен еще целый ряд глаголов. Они имеют общее значение ликвидации результата действия. Это глаголы *разбинтовать, развернуть, развязать, развьючить, разгрузить, расстегнуть* и т. п., описанные в книге [Апресян 1974: 147]. «У глаголов типа *забинтовать (руку марлей), заворачивать (покупку в бумагу), завязывать (ящик веревкой)* имеется валентность средства-объекта, нереализуемая у их антонимов *разбинтовать (руку), разворачивать (покупку), развязывать (ящик)*, хотя в принципе она у них есть, потому что их невозможно истолковать, не упоминая некоего Х-а (средства-объекта), который А снимает с В. Аналогичную семантическую структуру имеют и другие антонимы этого класса, к числу которых можно отнести, помимо уже названных слов, *развьючить (осла), разгрузить (машину), разжать (уши), расстегнуть (пальто)* и многие другие (ср. исходные глаголы *навьючить*

---

[3] Возможность высказываний типа *Из винтовки с оптическим прицелом и при хорошей видимости я по такой мишени не промахнусь* подробно обсуждается в книге [Мельчук 1974: 135] и в работе [Перцов 2006]. В связи с этим возникает вопрос: верно ли, что у глагола *промахнуться* нет синтаксического актанта на выражение «мишени». Обоснование разных точек зрения на эту тему даны в цитированных работах, а также в рецензии [Урысон 2008].

(осла тюками), нагрузить (машину зерном), зажать (уши ладонями), застегнуть (пальто на все пуговицы)» [Апресян 1974: 147]. У всех таких глаголов обсуждаемый актант может вообще не упоминаться в тексте. Ср. *Чем он зарабатывает? — Вагоны разгружает; Разверни, пожалуйста, этот пакет* и т. п.

Что касается обсуждаемых наречий, то выражение их актанта ‘ориентир’ в тексте совершенно обязательно. Высказывание типа *Вокруг стояли люди* или *Рядом росла береза* обязательно предполагает предтекст, в котором этот актант назван. Ситуация парадоксальна: семантический актант, с одной стороны, по определению «невыразим», а с другой стороны, его выражение в тексте совершенно обязательно.

Для представления рассмотренных фактов мы воспользуемся концепцией И. М. Богуславского, введя, однако, новые определения.

2.1. Разные типы «невыразимых» семантических актантов

Обычно семантический актант некоторой лексемы выражается в тексте по достаточно строгим регулярным правилам, имеющим синтаксическую природу; таковы, например, актанты глагола, выражаемые дополнениями. Однако выражение семантических актантов некоторых лексем не подчиняется подобным синтаксическим правилам; таков, например, актант ‘мишень’ глагола *промахнуться* или актант ‘ориентир’ наречия *вокруг*. Эти случаи требуется различать.

Примем, вслед за И. М. Богуславским, что синтаксический актант, точнее, синтаксическая сфера действия лексемы по некоторому семантическому актанту — это любое синтаксическое образование, значение которого соответствует семантическому актанту. Из изложенного ясно, что в языке представлено два типа синтаксических актантов (синтаксических сфер действия). Синтаксическая сфера действия (синтаксический актант) первого типа полностью описывается лингвистическими правилами. Будем называть такие синтаксические актанты лингвистически хорошо определенными. Синтаксические актанты второго типа не описываются лингвистическими правилами. Будем называть такие синтаксические актанты лингвистически неопределенными. Более точно:

- Если синтаксическая сфера действия лексемы по некоторому семантическому актанту вполне описывается строгими лингвистическими правилами, будем говорить, что данная синтаксическая сфера лингвистически хорошо определена.

Хорошо определены синтаксические сферы действия, соответствующие «классическим» синтаксическим актантам, например глагольным дополнениям. Хорошо определенными являются и синтаксические сферы действия частиц, описанные в книге [Богуславский 1996].

- Если синтаксическая сфера действия лексемы по некоторому семантическому актанту не может быть описана строгими лингвистическими правилами, будем говорить, что данная синтаксическая сфера является лингвистически неопределенной.

Примеры: синтаксическая сфера действия семантического актанта «мишень» глагола *промахнуться*; синтаксическая сфера действия семантического актанта «объект» глагола *разгружать*.

Среди лингвистически неопределенных синтаксических сфер действия выделяется важный подкласс. Это такие сферы действия, которые частично описываются правилами. Такова синтаксическая сфера действия наречий типа *вокруг* и *рядом* по семантическому актанту ‘ориентир’: про эту сферу действия известно, что она находится в предтексте относительно наречия.

Более точно:

- Будем говорить, что синтаксическая сфера действия лексемы по данному семантическому актанту является частично лингвистически определенной, если выражение данного актанта описывается формальными правилами не полностью.

Примеры других частично лингвистически определенных сфер действия приведены в работах [Урысон 2012; 2013].

Для нахождения семантического актанта, которому соответствует частично определенная синтаксическая сфера действия, требуются не только собственно лингвистические сведения (например, о линейном расположении фрагментов текста), но и понятийный анализ текста, невозможный без привлечения обширных «энциклопедических знаний» о действительности. Возможно, подобная процедура и поддается какой-то алгоритмизации, но при этом предполагается обработка не синтаксической или семантической структуры высказывания (текста), а некоторой понятийной сети.

3. Существуют ли наречные предлоги?

Выше мы уже говорили, что толковые словари представляют наречие и совпадающий с ним (в том числе и по значению) наречный предлог как две разные лексемы одного слова.

Недостаток этого подхода — в «умножении сущностей»: казалось бы, две в целом одинаковые единицы, различающиеся лишь синтаксически, вполне могут быть описаны в рамках одного значения. Так, в [Проспект 2010] принято, что некоторые лексемы имеют более одной модели управления, ср. *назначить его председателем — назначить председателя*; см. систематизацию случаев такого рода в работе [Апресян 2010]. Попробуем применить такой подход к нашему материалу.

При таком подходе слово *вокруг* описывается как наречие не только в контекстах типа *Вокруг стояли люди*, но и в контекстах типа *Вокруг костра стояли люди*, причем считается, что в обоих случаях представлена одна и та же лексема данного слова, имеющая, в частности, семантический актант ‘ориентир’. В синтаксической зоне словарной статьи наречия *вокруг* отмечается, что данный семантический актант может быть выражен двумя способами: (а)

существительным в предтексте; (б) формой родительного падежа существительного в постпозиции к данной лексеме (*вокруг дома*).

Аналогичным образом описываются и составные наречные предлоги. Например, наречие *рядом* признается наречием и в случаях типа *стоять рядом*, и в случаях типа *(стоять) рядом с кем/чем-л.* Во втором случае считается, что семантический актант 'ориентир' лексемы выражен предложно-падежным сочетанием *с кем/чем-л.*

(Мы сейчас отвлекаемся от случая, когда 'ориентиром' является говорящий или наблюдатель, ср. *Это рядом*.)

При таком подходе наречные предлоги вообще не выделяются. Достоинство этого подхода — не только в экономности, но и в более последовательном представлении некоторых фактов. Рассмотрим их.

Некоторые наречия, в частности пространственные, могут иметь зависимое наречие меры или степени; ср. *совсем рядом, довольно близко*. Эта способность иметь зависимое слово не связана с тем, как выражен семантический актант обсуждаемого наречия — падежной формой (предложно-падежным сочетанием) или существительным в любой форме в предтексте. Ср.

(6) а) *В парке был корт, а совсем рядом небольшая площадка с тренажерами.*
б) *В парке был корт, а совсем рядом с кортом небольшая площадка с тренажерами.*

(7) а) *Мы стояли и разговаривали, он стоял довольно близко и слушал.*
б) *Мы стояли и разговаривали, он стоял довольно близко от нас и слушал.*

Если считать, что в случаях (6б) и (7б) представлен наречный предлог, то придется признать, что предлог способен иметь зависимое наречие меры или степени. Это очень усложняет описание предлогов. Проще считать, что слова *рядом*, *близко* и т. п. во всех контекстах являются наречиями и сохраняют свойство иметь зависимое наречие.

У некоторых наречий обсуждаемый семантический актант выражается только одним способом — предложно-падежным сочетанием или падежной формой. Ср. наречия *задолго* и *незадолго*. Эти наречия имеют семантический актатнт 'временной ориентир', который может быть выражен только сочетанием *до чего-л.* Ср. *прийти задолго до начала спектакля, прийти незадолго до конца лекции*; при этом невозможно **Лекция начиналась в семь вечера, он пришел (не) задолго*. Если принять, что *задолго до* и *незадолго до* — это наречные предлоги, то останется неясным, какое наречие выделяется в их составе: такого наречия в языке просто не существует.

Мы привели аргументы против выделения группы наречных предлогов и в пользу того, чтобы все наречные предлоги считать наречиями. Рассмотрим теперь возможные аргументы в пользу описания, принятого в [Русская грамматика 1980].

Напомним, что в соответствии с этим описанием в случаях типа *Вокруг стояли люди* и *Вокруг костра стояли люди* представлены две разные лексемы слова *вокруг*, различающиеся своей частеречной принадлежностью (наречие vs предлог). Этим различием обусловлено и различие в синтаксических

свойствах данных лексем. Достоинство этого подхода состоит в четком разграничении лексических единиц по частеречному признаку. Действительно, и наречие, и предлог относятся к неизменяемым частям речи. Различаются они лишь синтаксически: наречие не обладает способностью управлять падежной формой существительного — это прерогатива предлога. Если отказаться от выделения наречных предлогов, то получится, что и предлог, и наречие могут управлять существительным, и тогда фактически стирается граница между наречием и предлогом как частями речи.

Однако многие наречия образованы от прилагательных, которые способны управлять. В [Русская грамматика 1980: 77] отмечается, что наречие наследует эту способность от производящего прилагательного. К этому классу управляющих наречий [Русская грамматика 1980] относит прежде всего предикативные наречия, т. е. наречия, которые (с некоторым упрощением) стоят в вершине предложения. Ср. *жаль кого-л. (Мне жаль его), стыдно за что-л. (Было стыдно за этот поступок), страшно за кого-л. (Ей страшно за детей)*. Обсуждаемые нами наречия не являются предикативными. Однако в данный класс в [Русская грамматика 1980: 77] попадают и некоторые из них, ср. *далекий <близкий> от чего-л. — далеко <близко> от чего-л.* Остается неясным, почему в этот класс не попали и другие наречные предлоги. С точки зрения словообразовательных связей между наречием и прилагательным все эти единицы (а не только *далеко от чего-л., близко от чего-л.*) естественно считать наречиями, а не предлогами.

Как видим, выделение наречных предлогов как подкласса предлогов не экономно и отчасти противоречит некоторым другим фрагментам грамматики.

На наш взгляд, в классе наречий естественно выделять подкласс управляющих наречий. Данный подкласс в свою очередь делится на две группы: управляющие предикативные наречия vs управляющие непредикативные наречия. Наречия из второй группы подобны предлогам: они управляют падежной формой слова или предложно-падежным сочетанием, но при этом и сами подчиняются какому-либо слову. Подчеркивая эту аналогию, можно называть данную группу слов «предложными наречиями». Однако нет достаточных оснований для того, чтобы называть эти слова «наречными предлогами» и, следовательно, помещать их в класс предлогов. Такое описание, как мы убедились, слишком громоздко.

Литература

1. *Апресян Ю. Д.* Лексическая семантика. М., 1974.
2. *Апресян Ю. Д.* Инструкция по составлению словарных статей Активного словаря (АС) русского языка // Проспект активного словаря русского языка / Отв. ред. Ю. Д. Апресян. М., 2010.
3. *Богуславская О. Ю.* Словарная статья синонимического ряда «БЛИЗКО 1.1» // Новый объяснительный словарь синонимов русского языка. Изд 2-е. / Отв. ред. Ю. Д. Апресян. М., 2004.

4. *Богуславская О. Ю.* Словарная статья синонимического ряда «БЛИЗКО 1.2» // Новый объяснительный словарь синонимов русского языка. Изд 2-е. / Отв. ред. Ю. Д. Апресян. М., 2004.
5. *Богуславская О. Ю.* Словарная статья синонимического ряда «БЛИЗКО 1.3» // Новый объяснительный словарь синонимов русского языка. Изд 2-е. / Отв. ред. Ю. Д. Апресян. М., 2004.
6. *Богуславский И. М.* Сфера действия лексических единиц. М., 1996.
7. *Мельчук И. А.* Опыт теории построения моделей «Смысл ↔ Текст». М., 1974.
8. НОСС — Новый объяснительный словарь синонимов русского языка. Изд 2-е. / Отв. ред. Ю. Д. Апресян. М., 2004.
9. *Перцов Н. В.* К суждениям о фактах русского языка в свете корпусных данных // «Русский язык в научном освещении», № 1 (11), 2006.
10. *Проспект* активного словаря русского языка / Отв. ред. Ю. Д. Апресян. М., 2010.
11. Русская грамматика, Т.1. М., 1980.
12. *Урысон Е. В.* Рец. на книгу: О. Н. Селиверстова. Труды по семантике. М., 2004. 959 с. // Известия РАН. Сер. лит. и яз. 2008, т. 67, № 3.
13. *Урысон Е. В.* Союзы, коннекторы и теория валентностей // Компьютерная лингвистика и интеллектуальные технологии. По материалам Международной конференции «Диалог» (2012). Вып. 11. Т. 1. М., 2012.
14. *Урысон Е. В.* Синтаксис союзов и коннекторов и теория валентностей // «Вопросы языкознания», № 3, 2013.

References

1. *Apresjan Yu. D.* Leksicheskaja semantika. M., 1974.
2. *Apresjan Yu. D.* Instrukcija po sostavleniju slovarnykh statej Aktivnogo slovarja (AS) russkogo jazyka // Prospekt aktivnogo slovarja ruisskogo jazyka / Otv. red. Apresjan Yu. D. M., 2010.
3. *Boguslavskaja O. Ju.* Slovarnaja stanja sinonimicheskogo riada “BLIZKO 1.1” // Novyj objasnitelnyj slovar’ sinonimov russkogo jazyka. Isd. 2-e. / Otv. red. Apresjan Yu. D. M., 2004.
4. *Boguslavskaja O. Ju.* Slovarnaja stanja sinonimicheskogo riada “BLIZKO 1.2” // Novyj objasnitelnyj slovar’ sinonimov russkogo jazyka. Isd. 2-e. / Otv. red. Apresjan Yu. D. M., 2004.
5. *Boguslavskaja O. Ju.* Slovarnaja stanja sinonimicheskogo riada “BLIZKO 1.3” // Novyj objasnitelnyj slovar’ sinonimov russkogo jazyka. Isd. 2-e. / Otv. red. Apresjan Yu. D. M., 2004.
6. *Boguslavskij I. M.* Sfera dejstvija leksicheskikh edinits. M., 1996.
7. *Melchuk I. A.* Opyt teorii postroenija modelej “SMYSL↔TEKST”. M., 1974.
8. *NOSS* — Novyj objasnitelnyj slovar’ sinonimov russkogo jazyka. Isd. 2-e. / Otv. red. Apresjan Yu. D. M., 2004.
9. *Pertsov N. V.* К суждениям о фактах русского языка в свете корпусных данных // «Русский язык в научном освещении», № 1 (11), 2006.

10. *Prospekt* aktivnogo slovarja ruisskogo jazyka / Otv. red. Apresjan Yu. D. M., 2010.
11. *Russkaja grammatika*, T. 1. M., 1980.
12. *Урысон Е. В.* Рец. на книгу: О. Н. Селиверстова. Труды по семантике. М., 2004. 959 с. // Известия РАН. Сер. лит. и яз. 2008, т. 67, № 3.
13. *Uryson E. V.* Sojuzy, konnektory i teorija valentnostej // Kompjuternaja lingvistika I intellektualnyje tekhnologii. Po materiailam Mezhdunarodnoj konferencii "DIALOG" (2012). Vyp. 11. T.1. M., 2012.
14. *Uryson E. V.* Sintaksis sojuzov i konnektorov i teorija valentnostej // "Voprosy jazykoznanija", № 3, 2013.

РЕГУЛЯРИЗАЦИЯ ВЕРОЯТНОСТНЫХ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ ДЛЯ ПОВЫШЕНИЯ ИНТЕРПРЕТИРУЕМОСТИ И ОПРЕДЕЛЕНИЯ ЧИСЛА ТЕМ[1]

Воронцов К. В. (voron@forecsys.ru)

Вычислительный центр им. А. А. Дородницына РАН;
Московский Физико-Технический Институт, Москва, Россия

Потапенко А. А. (anya_potapenko@mail.ru)

Московский Государственный Университет
им. М. В. Ломоносова, Москва, Россия

Вероятностное тематическое моделирование — это современный инструмент статистического анализа текстов, предназначенный для выявления тематики коллекций документов. Задача построения тематической модели имеет бесконечно много решений, что приводит к неустойчивости и плохой интерпретируемости тем. Для решения этих проблем применяется новый многокритериальный подход — аддитивная регуляризация тематических моделей (ARTM). Вводятся четыре регуляризатора: для выделения слов общей лексики в отдельные фоновые темы, для повышения разреженности и различности основных предметных тем, для удаления незначимых тем. В экспериментах показывается, что комбинирование этих регуляризаторов улучшает разреженность, когерентность, чистоту и контрастность тем без значимого ухудшения правдоподобия модели.

Ключевые слова: вероятностная тематическая модель, латентное размещение Дирихле, вероятностный латентный семантический анализ, регуляризация

REGULARIZATION OF PROBABILISTIC TOPIC MODELS TO IMPROVE INTERPRETABILITY AND DETERMINE THE NUMBER OF TOPICS

Vorontsov K. V. (voron@forecsys.ru)

Dorodnicyn Computing Centre of RAS;
Moscow Institute of Physics and Technology, Moscow, Russia

Potapenko A. A. (anya_potapenko@mail.ru)

Lomonosov Moscow State University, Moscow, Russia

[1] Работа выполнена при поддержке Российского фонда фундаментальных исследований, проекты 14-07-00847, 14-07-00908, 14-07-31176.

Probabilistic topic modeling is a rapidly developing branch of statistical text analysis. The topic model uncovers a hidden thematic structure of the text collection. Learning a topic model from a document collection has an infinite set of solutions. The nonuniqueness results in a weak interpretability and instability of the solution. To tackle these problems we use a new multiobjective approach — Additive Regularization of Topic Models (ARTM). ARTM is a non-Bayesian framework free of redundant probabilistic assumptions, which dramatically simplifies the inference of topic models and makes topic models easy to design, infer, and explain. With ARTM we combine four regularizers to concentrate common vocabulary words in background topics, to make domain topics sparse and distinct, and to eliminate insignificant topics. In our experiments the combination of the regularizers improves sparsity, coherence, purity, and contrast criteria at once almost without any loss of the perplexity.

Keywords: probabilistic topic modeling, Probabilistic Latent Semantic Analysis, Latent Dirichlet Allocation, regularization, EM-algorithm

1. Введение

Вероятностное тематическое моделирование (probabilistic topic modeling) — это современный мощный инструментарий статистического анализа текстов, предназначенный для выявления латентных тем в коллекциях документов [Blei 2012]. *Вероятностная тематическая модель* (ВТМ) определяет тему (topic) как совокупность слов, которые часто употребляются совместно в документах коллекции. Например, в коллекциях научных публикаций темы могут соответствовать явлениям, теориям, методам, при описании которых используется устоявшаяся терминология. В коллекциях новостных сообщений темы могут соответствовать событиям, странам, компаниям, персонам и т. д.

ВТМ осуществляет «мягкую» кластеризацию слов и документов по кластерам-темам. «Мягкость» означает, что слово или документ могут относиться к нескольким кластерам-темам с различными вероятностями. Тем самым выявляется тематическая структура документов, а также решаются проблемы синонимии и омонимии, возникающие при обычной «жёсткой» кластеризации. Синонимы, часто употребляющиеся в схожих контекстах, с большой вероятностью попадают в одну тему. Омонимы, употребляющиеся в разных контекстах, распределяются между несколькими темами соответственно частоте их употребления.

Вероятностный латентный семантический анализ (probabilistic latent semantic analysis) PLSA [Hofmann 1999] и *латентное размещение Дирихле* (latent Dirichlet allocation) LDA [Blei 2003] считаются стандартными методами вероятностного тематического моделирования и часто используются в прикладных исследованиях. В литературе описаны сотни их обобщений и модификаций [Daud 2010], имеются доступные реализации. Несмотря на интенсивный поток исследований в этой области, многие проблемы, в частности, проблемы неустойчивости и слабой интерпретируемости тем, пока не имеют окончательного решения.

Интерпретируемость тем является плохо формализуемым требованием. Предполагается, что, увидев список наиболее частотных слов и документов темы, человек сможет понять, о чём эта тема, дать ей адекватное название, определить более общие, более частные или близкие темы. Интерпретируемость тем важна для приложений тематического моделирования — информационного поиска, категоризации, аннотирования, сегментации текстов. Интерпретируемость тем позволяет систематизировать, визуализировать, объяснять результаты, выдаваемые пользователю информационной системы.

Однако темы, найденные с помощью ВТМ, часто оказываются непонятными, содержат слишком много слов, включают слова общей лексики, кажутся смесью нескольких слабо связанных тем, оказываются слишком похожими на другие темы. Более того, многократное обучение модели по одной и той же коллекции может давать совершенно разные темы в зависимости от случайного начального приближения. Исследователи либо мирятся с этими недостатками, не добиваясь понятности латентных тем, либо отказываются от применения ВТМ, не находя достойных альтернатив доступным реализациям PLSA и LDA.

Фундаментальная причина этих недостатков в том, что задача построения ВТМ по коллекции документов имеет бесконечно много решений, лишь малая доля которых интерпретируемы. Алгоритм оптимизации ВТМ выдаёт некоторое произвольное решение из этого множества.

Задачи, решение которых не существует, не единственно или не устойчиво, в математике принято называть *некорректно поставленными (по Адамару)*. Известен общий подход к их решению, называемый *регуляризацией*. Он заключается в том, что для выбора наилучшего решения задаются дополнительные критерии оптимальности, учитывающие специфические требования решаемой задачи и называемые *регуляризаторами*. Если вводится несколько критериев, то задача оптимизации становится многокритериальной. В данной работе рассматриваются требования интерпретируемости и предлагается их формализация в виде набора из четырёх регуляризаторов.

К сожалению, возможности гибкого введения регуляризаторов в PLSA и LDA не предусмотрены ни в теории, ни в реализациях. Современные ВТМ основаны на байесовском подходе, в котором комбинирование регуляризаторов вызывает структурные изменения модели и наталкивается на значительные технические трудности. Попытки построения многоцелевых ВТМ на основе LDA и генетических алгоритмов [Khalifa 2013] представляются громоздкими и вычислительно неэффективными. Большинство байесовских моделей, начиная с LDA, используют в качестве основного регуляризатора априорное распределение Дирихле. Это довольно сильное вероятностное допущение, которое не имеет убедительных лингвистических обоснований, не улучшает интерпретируемость и устойчивость модели, и затрудняет комбинирование регуляризаторов.

В данной работе для комбинирования регуляризаторов применяется новый подход, альтернативный байесовскому — *аддитивная регуляризация тематических моделей*, ARTM [Vorontsov 2014]. Он свободен от избыточных вероятностных допущений, не требует введения распределений Дирихле и позволяет использовать регуляризаторы, вообще не имеющие вероятностной

интерпретации. Включение ещё одного регуляризатора в модель выполняется «в одну строку» по готовым формулам, намного проще, чем в байесовском подходе.

Предлагаемый подход к повышению интерпретируемости основан на предположении, что если тема интерпретируема, то в ней имеется *ядро* — множество характерных слов, являющихся терминами определённой предметной области, которые с большой вероятностью употребляются в данной теме и практически не употребляются в других темах. Отсюда вытекают требования разреживания и повышения различности тем, переноса слов общей лексики в отдельные «фоновые» темы, и удаления незначимых тем. В данной работе эти требования формализуются с помощью комбинации четырёх регуляризаторов.

Большинство существующих методов оценивания интерпретируемости основано на привлечении экспертов-асессоров. В исследовании [Newman, 2009] экспертам предлагалось непосредственно оценивать полезность тем по трёхбалльной шкале. В методе word intrusion [Chang 2009] для каждой найденной темы составляется список из 10 наиболее частотных слов, в который внедряется одно случайное слово. Тема считается интерпретируемой, если подавляющее большинство экспертов правильно указывают лишнее слово. Экспертные подходы необходимы на стадии исследований, но затрудняют создание полностью автоматических технологий построения ВТМ. В серии работ [Newman 2009, Newman 2010, Mimno 2011] удалось найти величину, которая хорошо коррелирует с экспертными оценками интерпретируемости, и при этом вычисляется по коллекции автоматически. Это *когерентность* (coherence), оценивающая, насколько часто наиболее вероятные слова темы встречаются рядом в данной коллекции или в Википедии. Когерентность на сегодняшний день остается основной мерой интерпретируемости, вычисляемой автоматически.

В данной работе для оценивания тематической модели используются стандартные меры качества (*контрольная перплексия, когерентность*) и предлагаются новые меры интерпретируемости тем (*размер ядра, чистота и контрастность*), не требующие привлечения асессоров.

Эксперименты на коллекции англоязычных статей научной конференции NIPS показывают, что комбинирование регуляризаторов позволяет строить сильно разреженные модели с лучшими показателями интерпретируемости, без значимого ухудшения правдоподобия (перплексии) модели.

2. Тематическая модель PLSA

Вероятностная тематическая модель (ВТМ) описывает процесс пословного порождения документов. Пусть каждая тема t задана условным распределением $\phi_{wt} = p(w|t)$ на множестве всех слов w, каждый документ d задан условным распределением $\theta_{td} = p(t|d)$ на множестве всех тем. Моделируется процесс порождения коллекции. Для получения очередного слова сначала выбирается тема из распределения $\theta_{td} = p(t|d)$, затем выбирается само слово из распределения . В результате каждый документ d описывается распределением $\phi_{wt} = p(w|t)$.

$$p(w|d) = \sum_t p(w|t)p(t|d) = \sum_t \phi_{wt}\theta_{td} \quad (1)$$

Построение ВТМ является обратной задачей по отношению к описанному порождающему процессу. По заданной коллекции документов требуется найти параметры модели ϕ_{wt}, θ_{td} и определить число тем. Это задача стохастического матричного разложения. Заданную матрицу вероятностей слов в документах $F = \|p(w|d)\|$ требуется представить в виде произведения двух матриц меньших размеров — матрицы вероятностей слов в темах $\Phi = \|\phi_{wt}\|$ и матрицы вероятностей тем в документах $\Theta = \|\theta_{td}\|$. Данная задача является некорректно поставленной, поскольку, если пара матриц (Φ, Θ) является её решением, то пары матриц вида $(\Phi S, S^{-1}\Theta)$ при некоторых S также будут её решениями.

В данной работе принимается гипотеза «мешка слов» — предположение, что порядок слов не важен для определения тем документа. Коллекция задаётся частотами n_{dw} слов w в документах d. Заметим, что многие известные ВТМ отходят от гипотезы «мешка слов», учитывая порядок слов, синтаксис предложений, выделяя вместо слов коллокации или словосочетания. Поэтому нельзя говорить, что гипотеза «мешка слов» является ограничением для всех ВТМ.

Для оценивания параметров модели $\phi_{wt} = p(w|t)$ и $\theta_{td} = p(t|d)$ решается задача максимизации логарифмированного правдоподобия:

$$L(\Phi, \Theta) = \sum_d \sum_w n_{dw} \log \sum_t \phi_{wt}\theta_{td} \;\to \max_{\Phi,\Theta}.$$

Обычно для этого применяется итерационный процесс, называемый *EM-алгоритмом*. Его легко объяснить, не прибегая к строгим выкладкам. Процесс начинается со случайной инициализации параметров модели ϕ_{wt} и θ_{td}. Каждая итерация состоит из двух шагов, «Е» (expectation) и «М» (maximization).

На Е-шаге по формуле Байеса оценивается вероятность $p(t|d, w)$ и число вхождений n_{dwt} каждого слова w в каждый документ d, связанных с темой t:

$$n_{dwt} = n_{dw}p(t|d, w); \; p(t|d, w) = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}} \quad (2)$$

На М-шаге параметры ϕ_{wt} и θ_{td} вычисляются как частотные оценки соответствующих условных вероятностей. Значение ϕ_{wt} пропорционально числу раз n_{wt}, когда употребление слова w было связано с темой t. Значение θ_{td} пропорционально числу слов n_{dt} в документе d, относящихся к теме t:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \quad n_{wt} = \sum_d n_{dwt}, \quad n_t = \sum_{d,w} n_{dwt};$$

$$\theta_{td} = \frac{n_{dt}}{n_d}, \quad n_{dt} = \sum_w n_{dwt}, \quad n_d = \sum_{w,t} n_{dwt}.$$

Для краткости эти формулы записывают через знак пропорциональности $\propto$, позволяющий опускать нормировочный множитель:

$$\phi_{wt} \propto n_{wt}; \quad \theta_{td} \propto n_{dt}. \tag{3}$$

Известно, что основные алгоритмы обучения моделей PLSA и LDA можно рассматривать как EM-подобные алгоритмы [Asuncion 2009], различающиеся порядком применения формул E-шага (2) и M-шага (3), модификациями M-шага в результате регуляризации, способами распределения частоты n_{dw} по темам. Детали реализации и отличия этих алгоритмов обсуждаются в [Vorontsov 2013].

3. Аддитивная регуляризация тематической модели

Пусть наряду с правдоподобием $L(\Phi, \Theta)$ требуется максимизировать регуляризатор $R(\Phi, \Theta)$ зависящий от параметров модели. Будем максимизировать сумму двух критериев $L(\Phi, \Theta) + R(\Phi, \Theta)$. Решение данной задачи находится EM-алгоритмом с модифицированной формулой M-шага [Vorontsov 2014]:

$$\phi_{wt} \propto \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+; \quad \theta_{td} \propto \left(n_{dt} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)_+; \tag{4}$$

где $(x)_+ = \max\{x, 0\}$. К функции $R(\Phi, \Theta)$ предъявляется требование непрерывной дифференцируемости. Она может быть суммой нескольких регуляризаторов, взятых с весами, называемыми *коэффициентами регуляризации*. Таким образом, алгоритм многокритериальной оптимизации BTM, независимо от числа критериев, может быть получен из обычных EM-подобных алгоритмов PLSA или LDA простой заменой формулы M-шага.

Для повышения интерпретируемости тем будем опираться на предположение, что хорошо интерпретируемая тема должна иметь ядро, состоящее из терминов предметной области, отличающих её от других тем. Такие темы будем называть *предметными*. Слова общей лексики должны концентрироваться в отдельных *фоновых* темах. Формализуем эти гипотезы с помощью регуляризаторов.

Регуляризатор для разреживания предметных тем основан на предположении, что каждая предметная тема состоит из небольшого числа слов словаря, вероятности остальных слов в распределении ϕ_{wt} равны нулю. Предполагается также, что каждый документ относится к небольшому числу предметных тем, вероятности остальных тем в распределении θ_{td} равны нулю. Вводится регуляризатор, максимизирующий расстояние между распределениями ϕ_{wt} и распределением слов в коллекции β_w, а также между распределением θ_{td} и заданным распределением α_t на множестве предметных тем. Если в качестве расстояния между распределениями взять дивергенцию Кульбака-Лейблера, то формула M-шага (4) примет простой вид:

$$\phi_{wt} \propto (n_{wt} - \beta_0\beta_w)_+; \quad \theta_{td} \propto (n_{dt} - \alpha_0\alpha_t)_+,$$

где β_0, α_0 — коэффициенты регуляризации. Чем они больше, тем больше вероятностей ϕ_{wt} и θ_{td} обращаются в нуль на каждой итерации. Разреживание позволяет достигать 95–99% нулевых значений без значимого ухудшения

правдоподобия модели [Vorontsov 2013]. На ранних итерациях EM-алгоритма коэффициенты регуляризации лучше оставлять равными нулю, затем, по мере сходимости, постепенно увеличивать. Стратегия постепенного разреживания позволяет избежать преждевременного обнуления вероятностей.

Регуляризатор для сглаживания фоновых тем формализует требование, чтобы предметные темы не содержали слов общей лексики. Для описания этих слов в модель вводятся *фоновые* темы, распределения которых должны быть похожи на распределение слов во всей коллекции β_w. Регуляризатор минимизирует дивергенции Кульбака-Лейблера между распределениями ϕ_{wt} фоновых тем и распределением β_w. Кроме того, фоновые темы должны присутствовать в каждом документе. Поэтому вводится второй регуляризатор, минимизирующий дивергенции Кульбака-Лейблера между θ_{td} и α_t для фоновых тем t. В результате формула M-шага (4) даёт оценки параметров, аналогичные модели LDA:

$$\phi_{wt} \propto n_{wt} + \beta_0\beta_w; \quad \theta_{td} \propto n_{dt} + \alpha_0\alpha_t,$$

где β_0, α_0 — коэффициенты регуляризации. Эффектом данного регуляризатора является сглаживание (увеличение) малых значений параметров ϕ_{wt} и θ_{td} для фоновых тем за счёт незначительного уменьшения больших значений. Фоновые темы собирают слова общей лексики, стоп-слова и редкие слова, которые исключаются из предметных тем в результате разреживания. Сглаживание фоновых тем является обобщением робастных моделей [Potapenko 2012], в которых фактически использовалась только одна фоновая тема.

Отметим, что сглаживание и разреживание описываются общей формулой и отличаются только знаком параметров β_w, α_t. Это позволяет одновременно сглаживать фоновые темы и разреживать предметные. В байесовском подходе такая возможность до сих пор оставалась незамеченной из-за ограничения неотрицательности параметров распределения Дирихле. Модель LDA описывает только сглаживание; для построения разреженных моделей до сих пор приходилось использовать довольно сложные вероятностные конструкции.

Регуляризатор для декоррелирования тем. Ещё одно требование к предметным темам состоит в том, чтобы они как можно сильнее различались. Данное требование формализуется регуляризатором, который минимизирует сумму ковариаций между распределениями ϕ_{wt} и ϕ_{ws} для всех пар тем t, s [Tan 2010]. Формула регуляризованного M-шага (4) в этом случае принимает вид

$$\phi_{wt} \propto (n_{wt} - \tau\phi_{wt}(\phi_w - \phi_{wt}))_+; \quad \phi_w = \textstyle\sum_t \phi_{wt},$$

где τ — коэффициент регуляризации. Декоррелирование приводит к разреживанию тем и к более чёткому выделению ядер тем, состоящих из слов w с сильно доминирующей вероятностью $p(t|w)$. Декоррелирование, как и разреживание, хорошо сочетается со сглаживанием фоновых тем.

Регуляризатор для сокращения незначимых тем формализует требование, чтобы в модели не было тем, к которым относится слишком мало слов. Такие темы имеют маломощное ядро из редких слов. Чтобы исключать эти темы

из модели, вводится требование разреженности распределения тем во всей коллекции $p(t) = \sum_d \theta_{td} p(d)$. Регуляризатор максимизирует дивергенцию Кульбака-Лейблера между $p(t)$ и равномерным распределением. Формула регуляризованного М-шага (4) в этом случае принимает вид

$$\theta_{td} \propto \left(n_{dt} - \tau\theta_{td}\frac{n_d}{n_t}\right)_+.$$

где τ — коэффициент регуляризации. Согласно этой формуле, если число слов n_t, отнесённых к теме t во всей коллекции, мало, то вероятности этой темы понижаются для всех документов, вплоть до обнуления t-й строки матрицы Θ. Данный регуляризатор позволяет оптимизировать число тем, если начинать итерации с заведомо избыточного числа тем.

4. Оценки качества и интерпретируемости модели

Многокритериальная оптимизация требует также и многокритериального подхода к оцениванию качества ВТМ. При комбинировании регуляризаторов предлагается изменять коэффициенты регуляризации в ходе итерационного процесса и следить за изменениями различных показателей качества модели.

Перплексия является общепринятой мерой качества ВТМ. Она показывает, насколько хорошо модель (1) приближает наблюдаемые частоты появления слов в документах. Качество модели тем выше, чем меньше перплексия. Перплексия измеряется по контрольной выборке документов, не используемых для построения модели. Это позволяет избежать занижения оценки в результате переобучения.

Разреженность модели — доля нулевых значений среди параметров ϕ_{wt} и θ_{td}, только для предметных тем.

Число тем может уменьшаться при обнулении строк матрицы Θ в результате действия регуляризаторов разреживания или сокращения незначимых тем.

Доля фоновых слов $\frac{1}{n}\sum_{d,w} n_{dwt}$, где n — длина коллекции в словах. Принимает значения от 0 до 1. Значения, близкие к 1, свидетельствуют о вырождении тематической модели, например, в результате чрезмерного разреживания.

Когерентность темы определяется как средняя совместная встречаемость двух слов по всем парам k наиболее вероятных слов темы [Newman 2010, Mimno 2011]. Совместная встречаемость оценивается как *поточечная взаимная информация (PMI)* по документам, в которых встречаются оба слова. Число k в большинстве работ полагают равным 10. Для получения более глубоких оценок мы также вычисляем ещё две оценки когерентности: при $k = 100$ и по ядрам тем.

В данной работе предлагаются новые меры интерпретируемости тематической модели, не требующие асессорских оценок, как и когерентность. Будем относить слово w к *ядру* темы t, если $p(t|w) > \delta$. В наших экспериментах $\delta = 0.25$. Обозначим ядро темы t через W_t и определим три показателя интерпретируемости темы.

Размер ядра $|W_t|$ должен быть не слишком мал, но и не слишком велик. Ядра, содержащие ориентировочно от 20 до 200 слов, представляются адекватными.

Чистота темы $\sum_{w \in W_t} p(w|t)$ определяется как суммарная вероятность слов ядра. Принимает значения от 0 до 1; чем выше, тем лучше.

Контрастность темы $\frac{1}{|W_t|} \sum_{w \in W_t} p(t|w)$ равна средней вероятности встретить слова ядра именно в данной теме. Принимает значения от 0 до 1; чем выше, тем лучше. Показывает, насколько хорошо ядро темы отличает её от остальных тем.

Интерпретируемость модели тем лучше, чем выше когерентность, чистота и контрастность всех её тем. Поэтому мы определяем соответствующие показатели качества всей модели путём их усреднения по всем предметным темам.

5. Эксперименты

Исходные данные. Эксперименты проводились на коллекции NIPS, содержащей 1700 текстов статей научной конференции Neural Information Processing Systems на английском языке. Суммарная длина коллекции $2.3 \cdot 10^6$, объём словаря $1.3 \cdot 10^4$. Предварительная обработка текстов включала приведение к нижнему регистру, удаление пунктуации, удаление стоп-слов с помощью библиотеки BOW toolkit [McCallum 1996]. Во всех экспериментах общее число тем равно 100, для сглаженно-разреженных моделей среди них выделяется 90 предметных и 10 фоновых тем.

На рисунках 1 и 2 приведены зависимости показателей качества модели от номера итерации. На каждом рисунке сравниваются результаты модели PLSA и регуляризованной модели. Показатели качества выведены на четырёх графиках друг под другом с одинаковой горизонтальной осью. Верхний график: по левой оси перплексия, по правой — разреженности матриц параметров Φ, Θ. Второй график: по левой оси число тем, по правой — доля фоновых слов. Третий график: по левой оси размер ядра, по правой — контрастность и чистота. Нижний график: по левой оси когерентность ядра, по правой когерентности top-10 и top-100.

Такие графики предлагается использовать на этапе построения тематической модели для мониторинга показателей качества модели в ходе итерационного процесса. В частности, эти графики дают понимание, какие эффекты производит каждый регуляризатор в отдельности, как они взаимодействуют в комбинации, как выбирать стратегию изменения коэффициентов регуляризации. Не имея возможности привести здесь результаты всех протестированных комбинаций регуляризаторов, перечислим только основные выводы.

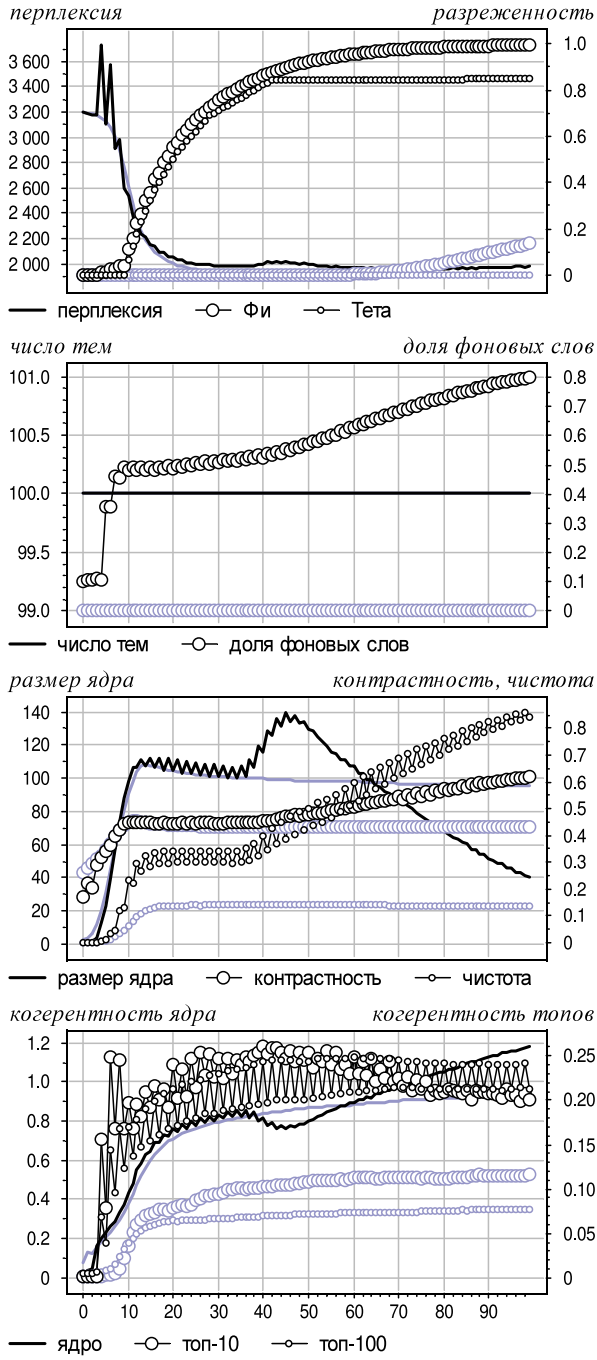


Рис. 1. Сравнение PLSA (серый) с комбинацией разреживания, сглаживания и декоррелирования (чёрный)

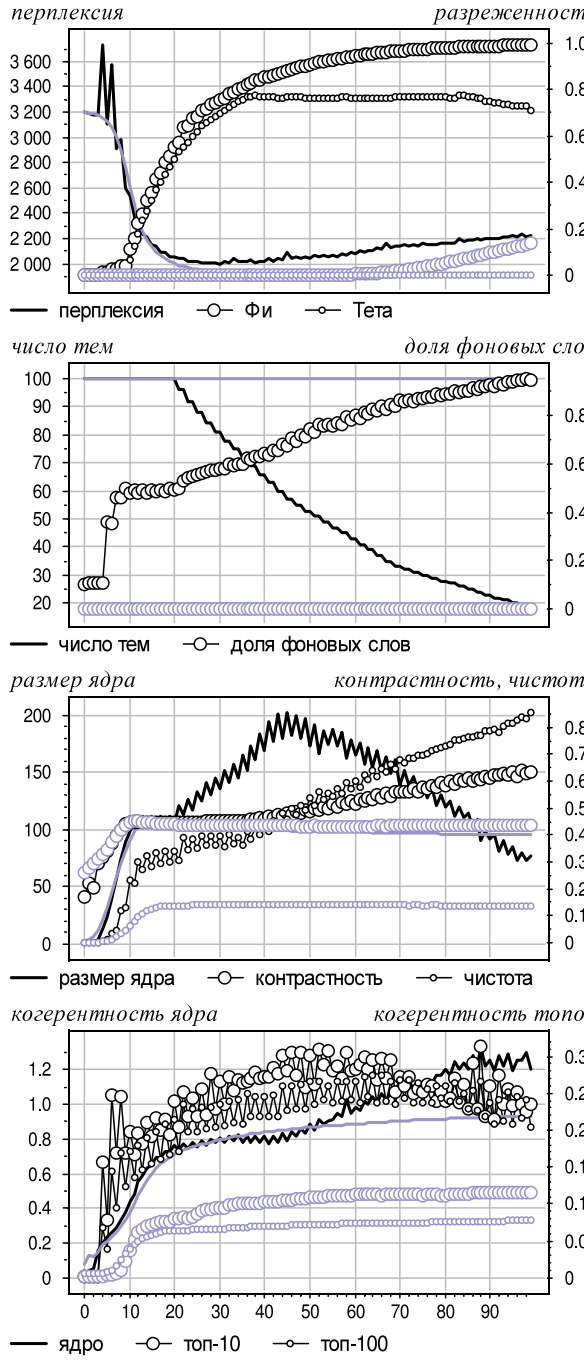


Рис. 2. Сравнение PLSA (серый) с комбинацией разреживания, сглаживания, декоррелирования и удаления тем (чёрный)

1. Разреживание предметных тем возможно до 98% в матрице и до 90% в матрице Θ практически без потери перплексии. Разреживание матрицы по равномерному распределению β_w приводит к удалению редких слов и улучшению контрастности. Если же в качестве β_w брать распределение слов в коллекции, то разреживание улучшает когерентность и чистоту. Коэффициенты регуляризации для разреживания рекомендуется плавно увеличивать после 10–15 итераций, когда итерационный процесс уже почти сошёлся, или хотя бы появилась определённость в том, какие именно элементы в матрицах Φ, Θ являются наименьшими.

2. Декоррелирование в несколько раз увеличивает чистоту и когерентность тем, но слабо разреживает матрицу Φ и вообще не разреживает матрицу Θ. Комбинация декоррелирования с разреживанием позволяет достичь сильной разреженности без уменьшения чистоты и когерентности. Декоррелирование рекомендуется включать с первой итерации, с максимально возможным коэффициентом регуляризации.

3. Сглаживание фоновых тем способствует переходу слов общей лексики из предметных тем в фоновые. Для этого достаточно одной фоновой темы. Сглаживание лучше включать с первой итерации, с фиксированным коэффициентом регуляризации. Комбинирование сглаживания фоновых тем с разреживанием и декорреляцией предметных тем достигает наилучших результатов по всей совокупности показателей (рис. 2).

4. Сокращение незначимых тем разреживает строки матрицы целиком, определяя минимальное необходимое число тем. Этот регуляризатор, так же, как и разреживающий, лучше включать постепенно, на фоне устойчивой сходимости процесса. Возрастание перплексии, уменьшение размера ядер или приближение доли фоновых слов к 1 могут свидетельствовать о вырождении тематической модели и нецелесообразности дальнейшего сокращения числа тем. В таком случае коэффициент регуляризации необходимо снова положить равным нулю. В наших экспериментах уменьшение числа тем ниже 60 ведёт к вырождению (рис. 2).

6. Выводы

Данная работа иллюстрирует применение нового подхода в тематическом моделировании, *аддитивной регуляризации тематических моделей* (ARTM), для построения сильно разреженной модели с интерпретируемыми темами. Предложены регуляризаторы разреживания предметных тем, сглаживания фоновых тем, декоррелирования и сокращения числа тем. Показано, что их комбинирование улучшает совокупность критериев качества практически без ухудшения перплексии модели. Предложена методика визуального мониторинга качества модели и подбора коэффициентов регуляризации. Поиск оптимальных *траекторий регуляризации* в пространстве коэффициентов регуляризации пока остаётся открытой проблемой.

Для оценивания интерпретируемости, наряду с когерентностью, предложены критерии чистоты и контрастности тем. Они основаны на гипотезе о том, что в интерпретируемой теме должно хорошо выделяться ядро — множество слов, отличающих данную тему от остальных.

Литература

1. *Asuncion A., Welling M., Smyth P., The Y. W.* (2009), On smoothing and inference for topic models, Proceedings of the International Conference on Uncertainty in Artificial Intelligence.
2. *Blei D. M., Ng A. Y., Jordan M. I.* (2003), Latent Dirichlet allocation, Journal of Machine Learning Research, Vol. 3, pp. 993–1022.
3. *Blei D. M.* (2012), Probabilistic topic models, Communications of the ACM, Vol. 55, No 4, Pp. 77–84.
4. *Chang J., Boyd-Graber J., Gerrish S., Wang C., Blei D. M.* (2009), Reading Tea Leaves: How Humans Interpret Topic Models, Advances in Neural Information Processing Systems, pp. 288–296.
5. *Daud A., Li J., Zhou L., Muhammad F.* (2010), Knowledge discovery through directed probabilistic topic models: a survey, Frontiers of Computer Science in China, Vol. 4, no. 2, pp. 280–301.
6. *Hofmann T.* (1999), Probabilistic latent semantic indexing, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA: ACM, pp. 50–57.
7. *Khalifa O., Corne D., Chantler M., Halley F.* (2013), Multi-objective topic modelling, Proceedings of 7th International Conference Evolutionary Multi-Criterion Optimization (EMO 2013), Springer LNCS, pp. 51–65.
8. *McCallum A. K.* (1996), Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering, http://www.cs.cmu.edu/_mccallum/bow
9. *Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A.* (2011), Optimizing semantic coherence in topic models, Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '11, pp 262–272.
10. *Newman D., Karimi S., Cavedon L.* (2009), External evaluation of topic models, Australasian Document Computing Symposium, December 2009, Pp. 11–18.
11. *Newman D., Lau J. H., Grieser K., Baldwin T.* (2010), Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '10, pp. 100–108.
12. *Potapenko A. A., Vorontsov K. V.* (2013), Robust PLSA performs better than LDA, 35th European Conference on Information Retrieval, Moscow, Russia, 24–27 March 2013, Lecture Notes in Computer Science, Springer Verlag-Germany, pp. 784–787.
13. *Tan Y., Ou Z.* (2010), Topic-weak-correlated latent Dirichlet allocation. 7th International Symposium Chinese Spoken Language Processing (ISCSLP), pp. 224–228.
14. *Vorontsov K. V., Potapenko A. A.* (2013), EM-like algorithms for probabilistic topic modeling [Modifikacii EM-algoritma dlja verojatnostnogo tematicheskogo modelirovanija], Machine Learning and Data Analysis [Mashinnoe obuchenie i analiz dannyh], Vol. 1, no. 6, pp. 657–686.
15. *Vorontsov K. V.* (2014), Additive Regularization for Topic Models of Text Collections, Doklady Akademii Nauk, Vol. 455, no. 3.

WHY STANDARD ORTHOGRAPHY? BUILDING THE USTYA RIVER BASIN CORPUS, AN ONLINE CORPUS OF A RUSSIAN DIALECT[1]

Waldenfels R. von (ruprecht.waldenfels@gmail.com)
Instytut Podstaw Informatyki Polskiej Akademii Nauk, Warsaw, Poland

Daniel M. (misha.daniel@gmail.com)
National Research University Higher School of Economics, Moscow, Russia

Dobrushina N. (nina.dobrushina@gmail.com)
National Research University Higher School of Economics, Moscow, Russia

The paper describes a corpus of dialectal Russian speech under development. The corpus relies on interviews conducted by a joint Swiss-Russian team in the summer of 2013 in a small cluster of North Russian villages with the goal of studying the local dialect from a sociolinguistic and dialectological perspective.
The interviews are transcribed into standard Russian and thus do not involve a detailed phonetic representation. The text is then lemmatized and grammatically annotated with standard tools and fed into a corpus. The corpus can be queried via a web-based interface which provides the user with access to the original sound recordings on a per-utterance level. This design, the paper argues, allows for a rapid development of the corpus without a major loss in usability, since the audio data are readily available. Future plans include more field trips as well as a more convenient interface providing, among other features, for user correction of the transcription.

Keywords: Russian; dialectology; corpus linguistics

1. Introduction

The abundance of linguistic data in corpora readily available over the internet has greatly changed the work of linguists in many subdisciplines. However, this

[1] This study (research grant No 14-05-0034) was supported by The National Research University–Higher School of Economics' Academic Fund Program in 2014. We thank the Slavic department of the University of Bern for hosting the corpus server.

development is probably least advanced in respect to the study of spoken language, and specifically corpora of dialectal or nonstandard speech. With some exceptions, corpus data for the study of dialects is still difficult to find on the web. Obviously, this is due to the specific challenges that spoken language poses in respect not only to collecting the data, but also to transcribing them to some written form, and making them available to the scientific community on the internet.

The present article describes a Russian dialect corpus project that aims to alleviate these problems by first transcribing the data into standard Russian, automatically annotating the corpus with standard tools, and making the result available on the internet together with aligned sound segments. This approach has a number of advantages, as well as some weaknesses.

The article is structured as follows. First we give a short overview of selected dialect corpora of Russian and other Slavic languages. We then introduce the place and circumstances of the Ustja River Basin Corpus data collection, before introducing the principles of and rationale for the transcription of the dialectal data in standard Russian. We then give an example analysis using the corpus data. Finally, we sketch planned further developments.

2. Corpora of Dialectal and Other Spoken Variants

Most Slavic dialect corpora make the data available in some sort of transcription that is usually situated between a faithful phonetic and a standard language representation; this is true, for example, for the Polish internet resource "Dialekty i gwary polskie" http://www.dialektologia.uw.edu.pl, for the dialectal and spoken data in the Czech National Corpus, the Slovak National Corpus, and others. The GOS corpus of spoken Slovene (http://www.korpus-gos.net/) offers both a standard and a more phonetically detailed transcription in two annotation layers. The Russian National Corpus contains a dialectal subcorpus which is mostly transcribed very near to the standard and is not very large (under 200,000 tokens), but offers the RNC's flexible search engine (powered by Yandex) for making sophisticated queries. The spoken subcorpus of the RNC is also orthography-oriented and uses a 'shallow' transcription, showing only pauses but not other discourse phenomena. The Saratov Dialect Corpus (http://www.sarteorlingv.narod.ru/projects.htm) offers detailed annotation as well as audio files.

Most of the above corpora do not include audio material. In some cases, as in the before mentioned "Dialekty i gwary polskie", the interviews are made available as full audio files alongside their transcription. In others, such as in the case of the German RuReg project (http://rureg.hs-bochum.de/), audio files of paragraph length are aligned to their transcription; the above mentioned Saratov Dialect Corpus seems to take a similar strategy (the corpus though was unavailable at the time of writing and submitting this paper).

Access to the original recording is especially important for dialect corpora, since transcription inevitably involves a loss in information that might be crucial for the analysis. We feel that for any extensive corpus based work on dialects or spoken data, it is crucial to have access to the audio files aligned to the transcribed text.

3. The Language of the Ustya River Basin: Collection and Corpus Composition

3.1. Data Collection and Transcription

The data of the Ustya Corpus was collected in the summer of 2013 by a joint group of Russian and Swiss students from the National Research University Higher School of Economics and the Slavic department of the University of Bern. The project is supported by HSE and carried out within the framework of a research and teaching cooperation between the two institutions.

The field team had its base in the village of Mikhalevskaya (locally known as Pushkino), in the Ustyan district of the Arkhangelsk Oblast, on the border with Vologda Oblast[1], but sometimes travelled a bit to the neighbouring villages. The students interviewed the villagers, asking people to tell them about their lives and other stories, and partly transcribed them on site, with more material transcribed later.

The dialects in Arkhangelsk Oblast have been object to vast research activities throughout the last century. This is why from the very beginning we intended to focus on studying variation rather than more preserved idiolects, modeling post-dialectal continuum rather than only the speech of oldest villagers (who are, as it happens today, mostly women). With some exceptions (e.g. Kochetov 2006, Krasovitsky 2013), the sociolinguistic dimension, mesolects and dialect attrition are still a rare topic in Russian dialectology.

The spoken data was transcribed using two programs: ELAN and Praat (since the formats are easily converted, transcribers were free to use either). In this type of transcription, each utterance in the audio file is marked and transcribed in one of several tiers. Informants are given separate tiers, with additional tiers for the interviewers, other speakers, and comments. The recorded speech is transcribed exclusively in standard Russian, with some provisions for marking unintelligible segments. The data is then stored in the ELAN XML format and processed further in an automatized procedure to add lemmatization and pos-tagging, and make it available over a web-based corpus interface.

Altogether, we collected some 40 hours of conversation. As of April 2014, 20 hours have been transcribed, comprising a corpus of around 215,000 tokens, of which about 180,000 tokens are informants' speech.

3.2. Why Standard Russian?

As indicated above, standard language is used, rather than a phonetic transcription as it is customary in most traditional publications. This means losing a lot of detail in comparison. Cf. the next texts from (Pozharitskaja 2005: 220, Vologda dialect), in both original transcription and the standard representation:

---

[1] We warmly thank our hosts Nikolaj Pushkin and Svetlana Pushkina for all their help in organizing our life, and work, and other practicalities.

[оп свojóй жы́з'н'е это хоц'ýгу погувур'и́т' / жы́с' мojá прошлá н'е óц'ен' ва́жно / жы́ла ф-так'и́jо го́ды т'ежóлыjо / д'ит'éй у м'ен'á бы́ло п'ét'еро / подн'алá ja д'ит'éй до войны́ / фторóй сын пог'и́п на войн'é]

"Об своей жизни это хочу поговорить. Жизнь моя прошла не очень важно, жила в такие годы тяжелые. Детей у меня было пятеро, подняла я детей до войны, второй сын погиб на войне."

Projects that adopt standardization approach are e.g., the Freiburg English Dialect Corpus (http://www2.anglistik.uni-freiburg.de/institut/lskortmann/FRED/), the ALCORP corpus of allemanic dialects of German, or the Nordic Dialect corpus (http://www.tekstlab.uio.no/nota/scandiasyn/). In many cases, the use of a standard language transcription is justified by focusing on (morpho)syntactic phenomena, which obviously does not assign high priority to phonetic detail (such is, for Russian, the dialectal subcorpus of the RNC). But in our case, standardization goes farther than in many other cases. At first glance, the corpus thus transcribed has nothing dialectal in it at all. What is its rationale?

In a nutshell, the standard orthography is justified by the fact that orthography is nothing but a key to the audiofiles with which the corpus is aligned. Corpora with standard transcription aligned with audio have been successfully used for phonetically oriented studies, e.g. by Streck (2012). The use of standard transcription means that we relegate a detailed phonetic analysis to a later stage (and probably to other experts). In general, this approach has the following advantages:

1. **Transcription into standard language can be done quickly.** When transcribing into standard orthography, there is no need to make difficult phonetic decisions concerning the data that involves repeatedly listening to the audio excerpt, comparing it to other segments of the main speaker, identifying phonetic variants—above all, a high expertise in dialectal phonetics. Note that even expert dialectologists may diverge on details of what they actually hear. While doing a standard transcription, it is sufficient to understand the text and identify the closest equivalent in the standard (however, below we discuss problems of defining what such an equivalent may be). This can be done much faster than phonetic transcription, and it demands by far less expertise.

2. **Transcription into the standard language effectively solves the problem of normalization and standardization.** Phonetic transcription systems used in different dialect corpora do not always coincide even for the same language, since the transcriber needs to balance readability and faithfulness to the sound shape, as well as decide what level of phonetic accuracy he or she wants to achieve for a given purpose. This is very difficult to do in a consistent way between transcribers, let alone different dialectal corpus projects. The standard language, in contrast, is well known to the transcribers and, in most cases, different transcribers will choose the same representation for a given dialect utterance without much doubt or need for consultation. This greatly reduces both systemic and non-systemic variation in the transcription of the same text by different transcribers.

3. **Transcription into the standard language makes the use of standard automatic annotation tools possible.** The automatic annotation of non-standard speech

is a difficult problem; see for example the system described in Wieczorek (2011) in the context of dialectal studies of Polish. Since we transcribe into standard Russian, however, we were able to use standard tools such as the TreeTagger (Schmid 1995) for the lemmatization and grammatical annotation of our data.

4. **Transcription into the standard language makes the data easily readable by non-linguist users.** In principle, the collected material may represent a cultural interest for a public broader than the dialectologists, including representatives of the local community, in the local towns if not in the village. Standard representation is much more suitable for the use of the interested 'lay' public, even if they are themselves speakers of the dialect (the combination of these two properties is however rare).

5. **Loss of phonetic data in transcription is made up for by aligning the transcription with the original audio.** Source audio information remains fully available to the user as the original audio is sentence-aligned to the transcription. Every user may make his or her own decision on what has been said, and how, and use examples from the corpus applying his or her own approach to dialectal transcription. For an expert, this is by far better than having to trust the transcriber.

4. An Overview of the Problems Related to the Transcription into Standard Language

The basic aim of the transcription is thus to provide the user with an easy access to the sound recordings. We do so by providing query interface based on standard automatic annotation tools. For this, the transcriber has to 'translate' or 'transpose' the dialectal text into standard language. This is far from being trivial, since many dialectal items on all linguistic levels do not have one-to-one correspondences in the standard. We will show several examples of such transpositions. Note that the transcription below (bracketed) is not intended to show the exact phonetic shape but to highlight the differences from the norm.

- If a dialectal word is different from the standard in a regular phonetic way, the standard variant is chosen: [заготовл'ёл'и] — *заготовляли*, [пр'ишóу] — *пришёл*, [пойи́с'] — *поесть*.
- Note that this includes cases where the standard correspondence may not be used in the sense in which it is used in the text; in such cases, we still use the standard word: [мои́ коро́вы шóбы не **рыч'ёл'и**] — *мои коровы чтобы не рычали* <чтобы мои коровы **не мычали**>.
- If a dialectal word is different from the standard in (the form of) the inflectional affix it takes, the standard variant is chosen: [р'евл'ý] — *реву*, [пок'исл'áje] — *покислее*, [мол'и́л'ис'е] — *молились*.
- A very frequently occurring phenomenon are postpositional particles, which correspond to the standard *-то*, but change their form depending on (the form of) the preceding word: [час'т'-ту] — *часть-то*, [тел'áта-та] — *телята-то*, [дом-от] — *дом-то*.

- If a dialectal word is different from the standard in the derivational affix it contains, the dialectal variant is chosen: [здал одно́ **кос'т'jо́**] — *сдал одно костьё* <такая худая была корова>, [бо́л'ше **н'е́как** бы́ло уч'и́ц'ц'е] — *больше некак было учиться* <больше никак не удавалось учиться>
- But if the difference in the derivational affix or in the root is trivial, the standard variant is used: [кварт'е́ра] — *квартира*, [робо́та] — *работа*, [топ'е́р'] — *теперь*.

The word 'trivial' here is not further formalized and relies very much on the intuition of the transcriber. All this boils down to the principle that, to make standard taggers applicable to the texts, we make as much phonetic adaptation as possible, reasonable and practicable without losing lexically, morphologically and syntactically relevant information—but not purely phonetic information which may be retrieved from the aligned audio. Thus, we certainly do not meddle with *meanings* and do not do *translation* of dialect texts into standards; and we do not force non-standard use of specific morphological forms into the rules of the standard language. Of course, this leaves us with many difficult cases when the transcriber has to make a decision that can hardly be formalized or generalized. For example, many dialects (including Ustja) use a standard word with different meaning—[немо́гу] corresponds to standard *не могу*, both phonetically (with an accent shift) and morphologically, but means 'to be ill'. One of our transcribers suggested that this verb should be written in the dialect as one word with the negation, as this combination has been clearly lexicalized and forms a new lexical item. We leave such subtle decisions to the transcribers, and assume that no exhaustive set of rules is possible or practicable. This may lead to some variation in transcription, but on the other hand will greatly facilitate transcribers' work.

As it is often the case in corpus building, we thus aim for a pragmatically sound, rather than for an ideal corpus, since going for such an ideal corpus would be certainly much more costly, perhaps in the end not feasible and quite conceivably unnecessary—i.e., a waste of resources. It is for this reason that in cases where several alternative solutions can be argued for and seem nearly equally plausible, we accept some variation between transcribers rather than try to achieve a completely consistent transcription and leave to the corpus user the task of dealing with potential divergences. This makes transcribers' task much more manageable, and may in fact in some cases lead to empirical solutions more robust than any theory-based inductive rule. Practice of corpus usage shows that many such theory based rules are non-intuitive anyway, and corpus users most often follow their intuitions rather than corpus descriptions. Note again that the fact that the raw data in form of audio segments are readily available means that any transcription can be checked by the user, making the transcription much less important than in traditional dialect texts.

4.1. Lemmatization, POS-Tagging and Inclusion into CWB

After transcription, the dialect text is lemmatized, tagged and imported to the Open IMS Corpus Workbench (CWB) corpus manager (http://cwb.sourceforge.net/) by a number of scripts, i.e., fully automatically once the transcribed file is entered into

the corpus repository. We use the TreeTagger (Schmid 1995) with a parameter file trained on the Multext-East tagset http://corpus.leeds.ac.uk/mocky/)

5. Using the Corpus

In this preliminary version, the data is accessible via a somewhat technical online interface that allows full CQP syntax as well as a simplified version of CQP (see Figure 1). Query results provide access to audio segments on an utterance level, so that the researcher has access to all properties that are lost under standardization—that is, to all phonetic, intonational or morphological details that are relevant for the research question the user is interested in. Expert users can check the correctness of the transcription (and, in the near future, will be able to add their comments to the texts of the corpus. Sample query results are shown in Figure 2.

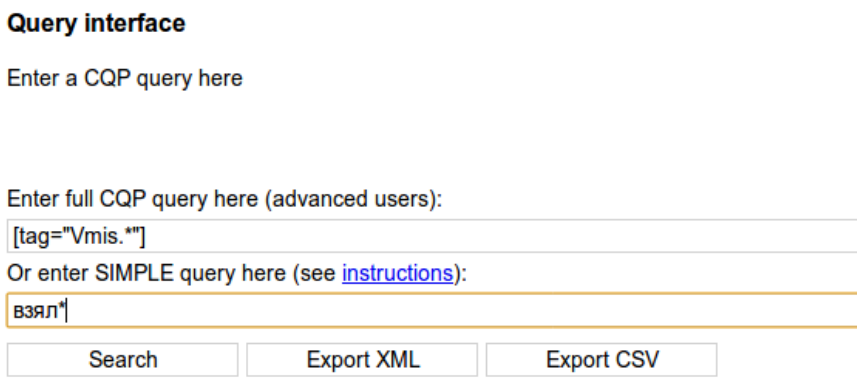


Fig. 1. Query interface

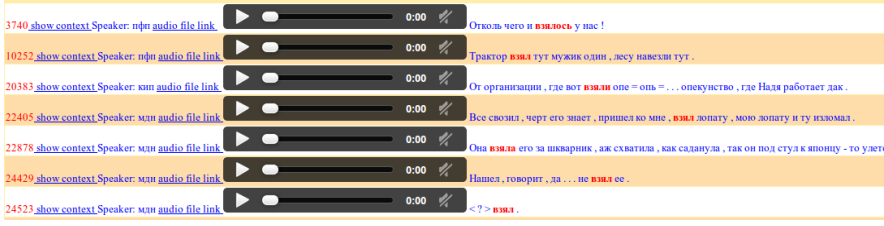


Fig. 2. Query results

The corpus interface provides access to lemmatization and grammatical information as parts of a query. For example, Seržant (2014) has used the corpus to investigate

the partitive genitive in Northern Russian. One way to find such partitive genitives is to use the query [tag="N.m.sg.*"] to look for all masculine genitive singular nouns in the standard transcription (the tagging uses the MULTEXT-EAST tagset for Russian, see http://corpus.leeds.ac.uk/mocky/msd-ru.html). The user then examines the audio files to decide which ending was used.

As a second example, consider the realization of /a/ as [e] between palatalized consonants that is found in many Russian dialects (Galinskaja 2005). To obtain all words where this change could have taken place (the envelope of variation), we look for all occurrences of "я" followed by one or more consonants, followed by a jotified vowel or soft sign in the standard transcription. We do so by using the regular expression query:

". *я([ртзпсдфгхклвбнм]*[ьеяиюё]|[цчжщй]). *"[2].

Using this query, we can quickly obtain the list of relevant word forms, which in the interviews with our oldest informant is as follows:

(30x) пятьдесят, (18x) пять, (13x) опять, (8x) объяснить, (6x) всякие, (5x) гуляет, прядет, (4x) прядешь, прясть, пятеро, пятисотку, (3x) девять, грязь, копятся, накопятся, память, представляете, пряли, тысяча, тысячи, (2x) Октябрьском, гулять, добавляли, доярки, завяжут, месяц, отправляли, отправляют, прялке, пяти, пятисотка, сеяли, сеялки, сплавляли, телятник, тысяч, тысячу, ячмень, (1x) Деревянние, Настоящая, Объяснишь, Отправляй, Прядешь, блядь, блять, вянет, гоняемся, граблями, грязе, грязи, гуляли, десять, дядя, завяжешь, заготовляли, заготовляют, запрягешь, заставляли, заставляют, кашляет, мягкие, напряде, напрядено, напрядет, настоящая, начислять, объявили, объясню, оставляют, отгоняли, поняли, пряди, прядёт, прялки, пятдесят, пятей, пятим, пятисотке, сеятся, справлялись, справляться, телятницей, телятся, трясти, тутошняя, тяжело, удобряют, яки, яме, ясли, яслях, ячмене

All the utterances with these words are displayed in the result window (cf. figure 2) and the users can examine and categorize these word forms in respect to the realization of /a/ as [e]. Preliminary analyses show that this change seems to be preserved only with older speakers; there is some evidence that it may be most resilient as a morphophonologically conditioned alternation in the language of younger speakers. But more research is necessary.

In sum, we see that for some questions, the representation in the standard transcription may be quite adequate (e.g., for some syntactic issues). For other questions, researchers need to do their own analysis of the audio data. In essence, thus, the painstaking work of a deep phonetic or other analysis is not performed in the transcription phase, as it is traditionally done, but at a later stage, and by the expert user him- or herself.

While this may seem as a drawback, note that since the analysis is done in the context of a specific research question, the accuracy of the analysis may actually be higher than in the context of a general-purpose phonetic transcription[3].

---

[2] Note that this expression is only an illustrative example that requires some later filtering, and does not cover /a/ before /j/, which is mostly lost in intervocalic surroundings.

[3] Of course, ultimately, it would be ideal if such annotations could be fed back into the corpus.

6. Further Plans

In 2014, the second field trip to Mikhalevskaya is planned, to make more recordings. The time in the field is also used as an opportunity for workshops on dialectal phonetics, morphology and syntax, for the students to exchange their ideas. As for the corpus, at this moment the interface is not yet publicly available on the web; access is granted only on an individual level. We are working on a more advanced interface that is more accessible to users that are not acquainted with the query language. Moreover, we want to enable users to correct mistakes in the transcription and in this way crowdsource some of the transcription work.

References

1. *Kochetov, Alexander* (2006). The role of social factors in the dynamics of sound change: A case study of a Russian dialect. Language Variation and Change, 18(01), 99–119.
2. *Krasovitsky, Alexander* (2013). Artikul'atsionnyj sdvig i razvitie nejtralizatsii glasnyh. In: British contributions for the XV Congress of Slavists, Minsk.
3. *Galinskaja, E. A.* (2005). Izmenenie [a] v [e] v istorii russkih dialektov. Vestnik Moskovskogo universiteta. Ser. 9, Filologija, (4), 42–54.
4. *Pozharitskaja, Sofia* (2005). Russkaja dialektologija. Moscow: Akademproekt.
5. *Schmid, Helmut* (1995): Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
6. *Seržant, Ilja* (2014) Independent partitive genitive in North Russian. In: Seržant, I. A. and B. Wiemer (eds.), Contemporary approaches to dialectology: The area of North, Northwest Russian and Belarusian vernaculars // Sovremennyje metody v dialektologii. Areal severnyh, severo-zapadnyh russkih i belorusskih govorov. Slavica Bergensia 13.
7. *Streck, Tobias* (2012) Phonologischer Wandel im Konsonantismus der alemannischen Dialekte Baden-Württembergs. Sprachatlasvergleich, Spontansprache und dialektometrische Studien. Stuttgart: Steiner (Zeitschrift für Dialektologie und Linguistik—Beihefte, Band 148).
8. *Wieczorek, Aleksandra* (2011). Słownictwo polskiej gwary kresowej na przykładzie Maćkowiec na Podolu. Charakterystyka funkcjonalna. Ph. D. Thesis, University of Warsaw.

ОЗВУЧИВАНИЕ ПИСЬМЕННОГО ТЕКСТА. КОРПУСНЫЙ И ИНСТРУМЕНТАЛЬНЫЙ АНАЛИЗ ПРОСОДИЧЕСКОЙ И КОММУНИКАТИВНОЙ СТРУКТУР ПРЕДЛОЖЕНИЯ[1]

Янко Т. Е. (tanya_yanko@list.ru)

Институт языкознания РАН, Москва, Россия

В работе анализируется коммуникативная структура предложений с начальным расположением группы, обозначающей новую информацию. В качестве точки отсчета используется анализ предложений с начальным новым, предложенный в работе [Ковтунова 1979] на примере текстов Пушкина и Л. Толстого. Русисты традиционно использовали для проверки своих научных гипотез тексты русских писателей. В настоящее время доступность текстов русской классики, озвученной лучшими носителями русской речи, и разработка современных компьютерных систем анализа устной речи, открывает доступ к интонационной структуре звучащего текста. Интонационная структура, в свою очередь, позволяет реконструировать коммуникативную структуру предложения, т. к. в устной речи интонация — это основное средство выражения коммуникативных значений. Для анализа создан исследовательский массив звучащих текстов русской классической литературы. Рассматриваются структуры, которые реально используются актерами, предположительно не читавшими работ И. И. Ковтуновой. В целом, корпусный и инструментальный анализ полностью подтверждает гипотезы И. И. Ковтуновой, сделанные на основе метода интроспекции, однако существенно расширяется спектр коммуникативных структур, возможных при реализации предложений с начальным новым.

Ключевые слова: коммуникативная структура, интонация, тема, рема, корпусные методы, чтение, звучащая речь, инструментальный анализ, данное, новое

[1] Работа над темой выполнена при поддержке РГНФ (грант 12-04-00258) и Президиума РАН по теме «Разработка компьютерной модели «Смысл — Звучащая речь» и электронной базы данных для ее поддержки» 2012–2014.

CORPUS AND INSTRUMENTAL METHODS IN ANALYSING FICTION AUDIO RECORDINGS

Yanko T. E. (tanya_yanko@list.ru)

Institute of Lingustics of the Russian Academy of Sciences, Moscow, Russia

This paper aims at analyzing the communicative structure of sentences with new information placed sentence-initially. The point of departure is the analysis of the examples from Pushkin and L. Tolstoy discussed in [Kovtunova 1979]. Russian linguists traditionally used Russian classical texts for verifying and exemplifying their scientific hypotheses. Currently, a vast amount of fiction read by the best actors and the development of convenient systems of spoken speech analysis, such as Praat or Speech Analyzer, open an easy access to the prosodic structure of a sentence. The prosodic structure, in its turn, allows of modelling the communicative structure, since prosody is the main means to manifest the communicative division of a sentence into theme and rheme. Availability of modern corpora and tools for investigation demonstrate the prosodic and the corresponding communicative structures really employed by the speakers who voiced over the texts of Pushkin and L. Tolstoy, and who presumably never read I. I. Kovtunova's papers. For investigation, a minor working corpus of sounding texts was set up. The new tools completely confirmed the basic I. I. Kovtunova's findings obtained in the second half of the 20th century by the method of introspection. Nevertheless, some new additions acquired by the use of the new sources of material and the corpus techniques correct the results achieved in [Kovtunova 1979] and substantially widen the variety of theme-rheme structures applicable to the sentences with new information placed sentence-initially.

Key-words: prosody, theme, rheme, given, new, instrumental analysis, corpus methods, reading, communicative structure

В работе анализируется коммуникативная структура предложений с начальным расположением группы, обозначающей новую информацию. В качестве точки отсчета используется анализ предложений с начальным новым, предложенный в работе [Ковтунова 1979] на примере текстов из Пушкина и Л. Толстого. Отталкиваясь от идей И. И. Ковтуновой, мы надеемся внести в ее результаты, полученные методом интроспекции, некоторые уточнения и дополнения, которые стали возможны благодаря доступности современных источников для анализа и компьютерных технологий.

Русисты традиционно использовали для проверки своих научных гипотез тексты русских писателей. В настоящее время произошел резкий сдвиг с анализа письменного языка классической литературы к анализу неподготовленной устной речи. В нашей работе мы предлагаем соединить две традиции — традицию

исследования письменного текста и традицию анализа звучащей речи — и вернуться к письменному тексту, проанализировав задачу его озвучивания. Совершить этот возврат нам позволит доступность текстов русской литературы, озвученной лучшими носителями русской речи. Таким образом, в данной работе ставится задача анализа особого типа звучащей речи: актерского чтения. Одно из преимуществ такого анализа состоит в том, что исследователь получает возможность сравнить стратегии чтения одного и того же текста разными исполнителями. При совпадении стратегий возникает гипотеза о том, что соответствующая просодическая и, соответственно, коммуникативная структура могли входить в замысел автора текста. Мы обратимся к вопросу о том, какие коммуникативные структуры реально используют в чтении Пушкина и Л. Толстого И. Смоктуновский и О. Табаков, предположительно не читавшие работ И. И. Ковтуновой.

Реконструкция коммуникативных структур, реализованных в чтении, становится возможной благодаря анализу интонационной структуры, использованной чтецами, т. к. в звучащей речи интонационная структура служит основным средством выражения членения предложения на тему и рему. Тема манифестируется подъемом тона на ударном слоге акцентоносителя темы, а рема — падением. Специфические коммуникативно релевантные изменения частоты основного тона отличают тему от ремы, а границы темы и ремы — манифестируются способом выбора словоформы-носителя акцентного пика в теме или в реме. Мы не можем осветить проблему выбора акцентоносителя темы или ремы, оставаясь в рамках короткой статьи, поэтому отсылаем читателя к решению, приведенному в [Янко 2008b, 38–60].

Теоретически анализ актерского чтения был доступен и во второй половине 20 века, однако легкого доступа к большим массивам записей, сделанных разными исполнителями, и удобных машинных систем анализа устной речи, таких, как Praat и Speech Analyzer, в то время не было. Технологии верификации перцептивных ощущений слушающего по данным приборов находились в руках специалистов по фонетике, и практика использования звучащих данных и инструментальных технологий специалистами по семантике и прагматике в то время еще не сформировалась. В настоящее время имеется эмпирическая база и технологии работы с ней, что создает основу для соединения двух линий исследования: теоретической и экспериментальной. На фоне традиционной установки русистов на анализ литературного письменного текста, с одной стороны, и направленности исследований последних лет на анализ, наоборот, неподготовленной устной речи — с другой, остается место и для третьей линии — анализа озвученного литературного текста, задачи, которая, с использованием современных средств анализа, насколько нам известно, конкретно раньше никем не ставилась.

Дальнейший план состоит в следующем. Исходной точкой анализа станут примеры из Пушкина и Л. Толстого и то членение предложения, по И. И. Ковтуновой, которое они иллюстрируют. Далее рассматривается актерское чтение этих примеров. Интонация чтения эксплицируется тональными кривыми, полученными с помощью системы анализа устной речи Speech Analyzer. Параллельно с примерами И. И. Ковтуновой на основе описания, которое их сопровождает, рассматриваются и другие, структурно идентичные

исходным предложения из русской литературы. Наблюдаемые интонационные структуры интерпретируются с точки зрения коммуникативного членения, которое ими манифестируется.

Рассматриваемый И. И. Ковтуновой особый тип предложений характеризуется начальным расположением именной группы, соответствующей новой2 информации: … *белая собачка английской породы залаяла и побежала ей навстречу* (Пушкин). (Группа, обозначающая новое, выделена разрядкой.) Предложения с препозицией нового дают интересный материал для анализа, т. к. заключают в себе отличную от базовой схему линейного расположения информационных компонентов русского предложения. Базовые предложения строятся по принципу «новое в конце»: *Ей навстречу побежала маленькая собачка*. Препозиция нового ставит перед исполнителем выбор интерпретаций: темой будет препозитивная группа или компонентом ремы? Этот вопрос возникает потому, что в соответствии с базовыми приоритетами русского языка при порождении сообщения сегмент, соответствующий новому, воплощается в рему, получает соответствующий реме интонационный показатель и занимает место ремы в исходе предложения. Между тем в анализируемых предложениях с порядком слов, возникшим под пером Пушкина и Л. Толстого, новое занимает линейное место не ремы, а темы. В результате, возникает вопрос, какой фактор «перевесит» при чтении: если информационный, то новое воплотится в рему, и она тем самым окажется в несвойственном реме месте — начале предложения, если возобладает фактор линейного расположения в предложении темы и ремы, то новое воплотится в тему с несвойственной ей референцией к новой информации. Записи чтения говорят о том, какой выбор делают актеры.

Для реализации плана автором был создан исследовательский массив записей актерского исполнения «Капитанской дочки», «Анны Карениной» и ранних рассказов И. Бабеля. Состав массива объясняется тем, что из «Капитанской дочки» и «Анны Карениной» были взяты примеры И. И. Ковтуновой. Для текстов Пушкина и Л. Толстого расположение нового в начале предложения вообще весьма характерно. Рассказы И. Бабеля также используются для анализа, т. к. в них расположение нового в начале стало специфическим, даже навязчивым, приемом построения текста. В корпусе имеются записи чтения одного и того же текста разными чтецами.

В разделе 1 ниже примеры И. И. Ковтуновой [1979] рассматриваются на конкретном материале актерского чтения. В разделе 2 анализируются другие примеры, имеющиеся в массиве, структурно идентичные исходным. В дополнение к анализу художественной литературы в разделе 3 рассматривается сложившаяся при реализации предложений, также построенных по принципу «новое в начале», традиция чтения текстов дикторами радио и телевидения.

---

[2] О противопоставлении данного и нового в информационной структуре предложения см. [Chafe 1976].

1. Интонационная реализация примеров из «Капитанской дочки» и «Анны Карениной»

И. И. Ковтунова пишет: «... в художественной прозе возможны... принципы построения текста, связанные с нестандартными способами введения новой информации… одним из таких принципов является включение новой информации… непосредственно в тему высказывания… Наиболее наглядно этот принцип обнаруживает себя в предложениях, в которых темой служит состав подлежащего, а ремой — состав сказуемого…: *Страшная буря рвалась и свистела между колесами вагонов по столбам из-за угла станции* (Л. Толстой). Приведенное предложение является началом главы и заключает в себе по существу два сообщения: 1) Была страшная буря; 2) Эта буря рвалась и свистела... В логически развернутом изложении новый предмет или явление, выраженное субстантивной группой, обычно вводится в контекст нерасчлененным высказыванием с экзистенциальным глаголом: *Была страшная буря* ... В последующих высказываниях даются характеристики этого явления. Но в художественном повествовании часто происходит сжатие двух сообщений в одно… Ср. другие примеры: …*Марья Ивановна пошла около прекрасного луга, где только что поставлен был памятник в честь недавних побед графа Петра Александровича Румянцева. Вдруг белая собачка английской породы залаяла и побежала ей навстречу* (А. Пушкин)» [Ковтунова, 1979, с. 263].

Итог этой трактовки такой. 1) Предложения с начальным новым имеют коммуникативную структуру Тема-Рема. 2) С семантико-прагматической точки зрения в этих предложениях заключено два сообщения: одно, вводящее в рассмотрение новый предмет или событие (‘Была буря’, ‘Появилась собачка’), и другое, характеризующее этот предмет с той или иной точки зрения (‘Эта буря рвалась и свистела’, ‘Собачка залаяла и побежала навстречу Марье Ивановне’).

Предложение (1) из «Анны Карениной» актеры О. Табаков и В. Герасимов единодушно интерпретируют как тему, дополнительно осложненную значением эмфазы.

(1) *Страшная буря рвалась и свистела между колесами вагонов по столбам из-за угла станции.*

Тонограмма[3] чтения примера (1) О. Табаковым на рисунке 1 говорит о том, что начальная группа *страшная буря* интерпретируется как эмфатическая тема, ударный слог акцентоносителя — словоформы *буря* — на тонограмме выделен овалом. О. Табаков читает это предложение в усеченном виде, что не имеет принципиального значения, т. к. основным объектом анализа здесь служит начальная группа.

---

[3] Тонограммы получены с помощью компьютерной системы Speech Analyzer. На графике по оси абсцисс откладывается время в секундах, по оси ординат — частота основного тона голоса в герцах.

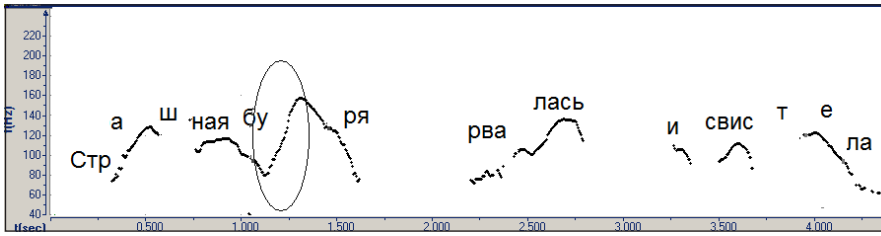


Рис. 1. Тонограмма примера (1) в исполнении О. Табакова.

Основное движение тона здесь — восходящее. Оно служит манифестантом темы. Кроме того, наблюдается предшествующее основному движению тона в противоположную сторону, которое «искривляет» подъем, что говорит об эмфатическом выделении4. То, что тема реализуется в эмфатической модификации, объясняется прямым указанием на то, что буря была страшная. На акцентоносителе конечной группы словоформе *свистела* наблюдается нисходящий акцент ремы. Перед нами коммуникативная структура Эмфатическая тема—Рема с эмфатической темой *страшная буря* и ремой *рвалась и свистела*.

В исполнении чтеца В. Герасимова этот пример дается в полной форме, как у Л. Толстого:

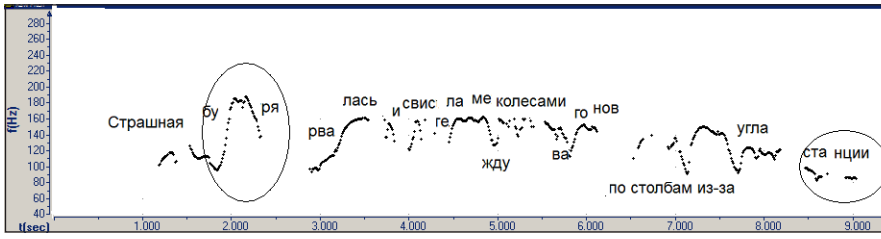


Рис. 2. Тонограмма примера (1) в исполнении В. Герасимова

Тонограмма демонстрирует подъем, манифестирующий тему, опять же в сопровождении эмфатического «искривления» на акцентоносителе темы словоформе *буря*, и сглаженное конечное рематическое падение на акцентоносителе ремы словоформе *станции*. (Слова-акцентоносители выделены овалами.) Трактовка В. Герасимова практически совпадает с трактовкой О. Табакова. Напомним, что И. И. Ковтунова предполагала в данном предложении как вариант для реализации группы *страшная буря* тему.

Обратимся к тонограммам чтения примера (2) из «Капитанской дочки» Н. Мартоном и И. Смоктуновским.

(2) *Вдруг белая собачка английской породы*
залаяла и побежала ей навстречу.

---

4 Более детально о семантике и средствах выражения эмфазы см. [Янко, 2008а].

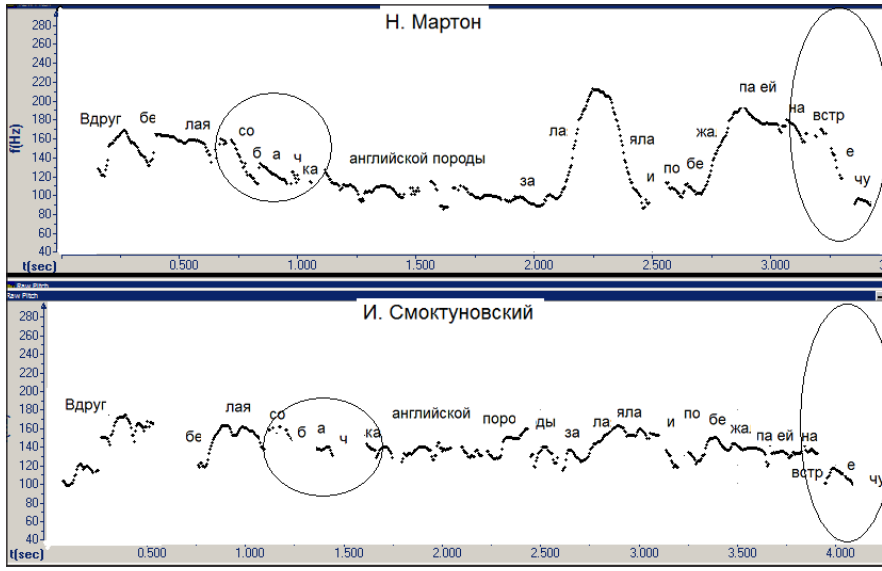


Рис. 3. Сравнительная тонограмма чтения примера (2) Н. Мартоном и И. Смоктуновским

Тонограммы чтения предложения (2) Н. Мартоном и И. Смоктуновским на рисунке 3 говорят о единодушной трактовке этого примера обоими актерами. Можно наблюдать падение — более рельефное у Н. Мартона на верхней панели и почти ровный тон (отличный, однако, от восходящего акцента темы, который тоже можно было бы здесь ожидать) у И. Смоктуновского на нижней панели — на словоформе *собачка* (на обоих графиках словоформа *собачка* выделена овалом)5 и второе падение на акцентоносителе второй ремы *навстречу* (тоже выделено овалом), опять же — более крутое у Н. Мартона и более пологое у И. Смоктуновского. Первая рема соответствует компоненту появления на сцене ‘Появилась собачка’, вторая — характеризующему ‘Она залаяла и побежала…’. В трактовке других, кроме рем, компонентов примера (2) тоже имеется существенное сходство: и у Н. Мартона, и у И. Смоктуновского наблюдается подъем, служащий просодическим коррелятом темы на *вдруг*, подъемы незавершенности на *залаяла* и *побежала*. Восходящие акценты и вносимые ими значения — это ожидаемый элемент трактовки, они не имеют здесь принципиального значения. Существенно, что препозитивная группа *белая собачка* получает в обоих чтениях акцент ремы, носителем которого служит словоформа *собачка*. Оба исполнителя интерпретируют пример из Пушкина как структуру с двойной ремой. В принципе, трактовка этого предложения могла быть и другой, например, интерпретацией с начальной темой, как полагала И. И. Ковтунова. Между тем в анализируемых вариантах чтения интерпретации совпали.

---

5 О сдвиге акцента в атрибутивных именных группах типа собачка английской породы с словоформы породы на словоформу собачка см. [Янко 2012].

Структура с двойной ремой наиболее точно отвечает семантической трактовке предложений с начальным новым о совмещении в одном предложении двух сообщений, потому что при двойной реме каждому сообщению соответствует отдельная рема. Так, в предложении (2) имеется тема *вдруг*, начальная рема *белая собачка английской породы* и конечная рема *залаяла и побежала ей навстречу.*

Пример (3) из «Капитанской дочки» стал доступен нам в чтении трех исполнителей: И. Смоктуновского, Н. Мартона и В. Самойлова.

(3) *Неожиданные происшествия, имевшие важное влияние на всю мою жизнь, дали вдруг моей душе сильное и благое потрясение.*

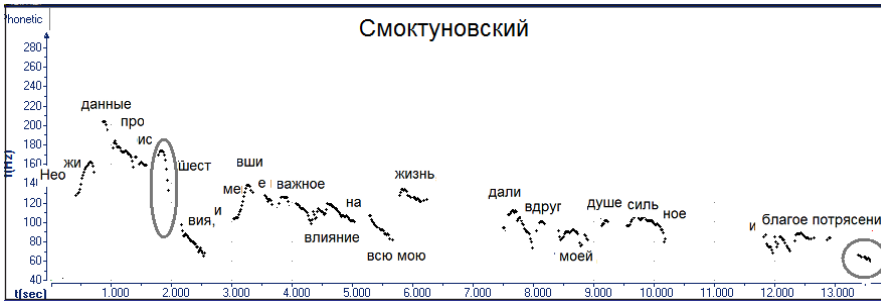


Рис. 4. Тонограмма примера (3). Чтение И. Смоктуновского

В чтении И. Смоктуновского ударный слог акцентоносителя начальной группы *неожиданные происшествия* словоформа *происшествия* получает нисходящее движение тона, которое характеризуется существенным перепадом частот (ударный слог выделен овалом). *Неожиданные происшествия* — это начальная рема, сегмент *дали вдруг моей душе сильное и благое потрясение* — конечная рема. Кроме того, в (3) имеется определение *имевшие важное влияние на всю мою жизнь*, которое получает естественную в данном случае коммуникативную и интонационную интерпретацию: этот сегмент расположен после акцентоносителя начальной ремы и несет соответствующий относительно ровный и низкий тон. Поскольку это определение содержит референцию к первому лицу (ср. местоимение *мою*), т. е. соотносится с известной информацией (данным), можно предположить, что и автор, и чтец трактуют это определение как заударную тему. Кроме того, финал этого сегмента словоформа *жизнь* несет указание на то, что предложение еще не кончилось: словоформа *жизнь* получает подъем незавершенности.

Рассмотрим чтение того же предложения Н. Мартоном. С точки зрения тема-рематической структуры Н. Мартон дает членение, которое фактически совпадает с членением И. Смоктуновского:

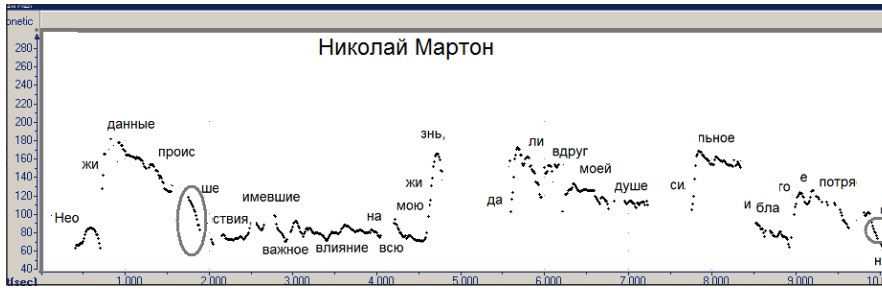


Рис. 5. Тонограмма примера (3). Чтение Н. Мартона

Наблюдается то же падение на *происшествия*, которое маркирует начальную рему, подъем незавершенности на *жизнь* и конечное падение на словоформе *потрясение*, которая служит акцентоносителем второй — характеризующей — ремы. И. Смоктуновский и Н. Мартон практически единодушны в просодической и, соответственно, тема-рематической интерпретации предложения (3). Интерпретация группы *неожиданные происшествия* как ремы поддержана семантикой словоформы *неожиданные*, которая сигнализирует об отсутствии объекта референции в зоне внимания слушающего. Это больше согласуется с функцией ремы, чем темы.

Третий исполнитель — В. Самойлов — дает тексту Пушкина иную интерпретацию.

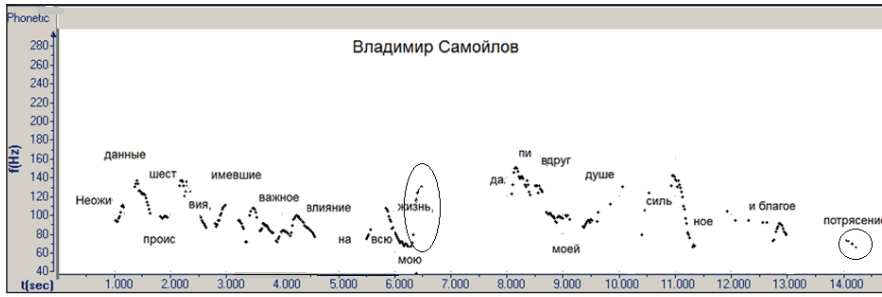


Рис. 6. Тонограмма примера (3). Чтение В. Самойлова

В. Самойлов интонирует начальную именную группу как тему, однако в состав темы он включает не только вершинную группу *неожиданные происшествия*, но и определение *имевшие важное влияние на всю мою жизнь*. Поскольку определение содержит референцию к первому лицу, вся именная группа, включая определение, может быть понята как известная адресату. Соответственно, группа *неожиданные происшествия, имевшие важное влияние на всю мою жизнь* реализуется как цельная тема с акцентоносителем словоформой *жизнь*. Акцентоноситель несет подъем тона, а остаток отходит к реме, о чем говорит падение на словоформе *потрясение*. В. Самойлов интерпретирует пример (3) как структуру Тема—Рема.

Анализ артикуляции этих и других примеров говорит о том, что наиболее частотным и эффектным приемом введения в рассмотрение нового объекта, которое одновременно сопровождается его характеристикой, служит структура с двумя ремами: начальной и конечной. Использование этой структуры — сугубо литературный прием. Он применяется в актерском чтении предложений, где начальная группа соотносится с новой информацией. К такой интерпретации чтеца подводит автор текста, располагающий новое в начале. В неподготовленной речи концентрация двух рем в одном речевом акте не используется. Предложениями, реализующими двойную рему, не говорят, а пишут, и двойная рема возникает в процессе чтения письменного текста. Кроме актерского чтения, коммуникативная структура с двумя ремами, заключенными в одном речевом акте, встречается в речи профессиональных лекторов и дикторов средств массовой коммуникации. На реализацию структуры с двумя ремами чтеца толкает не только расположение нового в начале, прием, который вводит в рассмотрение новый бытующий предмет, но также и достаточно весомый характеризующий компонент, расположенный «справа». При отсутствии характеризующего компонента может быть реализована структура с одной начальной ремой; примеры за недостаточностью места мы опускаем. Кроме того, «рематической» реализации с начальной ремой способствует компонент новизны, неожиданности, странности в семантике начальной группы, ср. лексему *неожиданные* в примере (3). Слова же, выражающие сильные чувства говорящего, ведут к реализации эмфатических модификаций начальной группы-нового, как виде темы, так и в виде ремы, ср. чтение лексемы *страшная* в примере (1). Примеры с начальной эмфатической ремой у нас имеются, но за ограниченностью места мы их также опускаем.

2. Некоторые другие коммуникативные трактовки предложений с новым в начале

Рассмотрим несколько коммуникативных реализаций предложений с новым в начале, которые не охватил предыдущий раздел. В чтении примера (4) из «Капитанской дочки» И. Смоктуновским начальная группа-новое реализуется как простая тема.

(4) *<Мысль о скорой разлуке со мною так поразила матушку, что она уронила ложку в кастрюльку,>*[6] *и с л е з ы потекли по ее лицу.*

[6] Необходимый контекст анализируемых примеров заключен в угловые скобки.

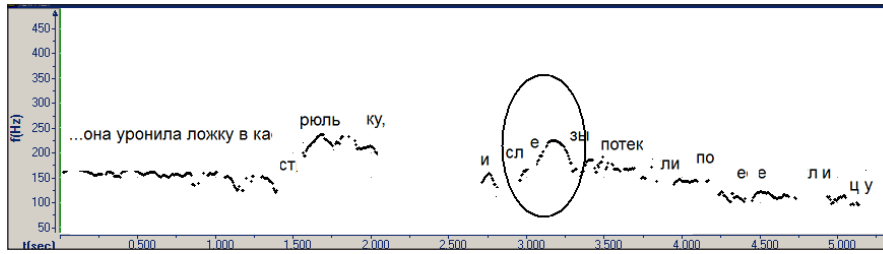


Рис. 7. Тонограмма примера (4)

«Тематическая» интерпретация начального нового объясняется тем, что слезы на лице у матери, которая думает о разлуке с сыном, трактуются исполнителем как ожидаемая реакция, или данное. Тонограмма демонстрирует подъем на ударном слоге акцентоносителя темы словоформе *слезы* плюс падение на заударном слоге. Это типичный акцент темы, ИК-3, по Е. А. Брызгуновой [1982, 103–118]. Словоформа *слезы* на тонограмме выделена овалом. Соответственно, *потекли по ее лицу* — это рема.

Предложения с начальным новым могут также изображать фон для событий, продвигающих повествование вперед. Мы называем здесь эту структуру «фоновой» ремой, хотя она имеет более широкий диапазон функций, чем маркирование фоновых событий. В примере (5) из рассказа И. Бабеля «Первая любовь» изложению действий героев противопоставлено описание интерьера, где разворачиваются события:

(5) *<Она обняла меня и повела по коридору … Мы пришли на кухню…> Гусь жарился на кафельной плите…*

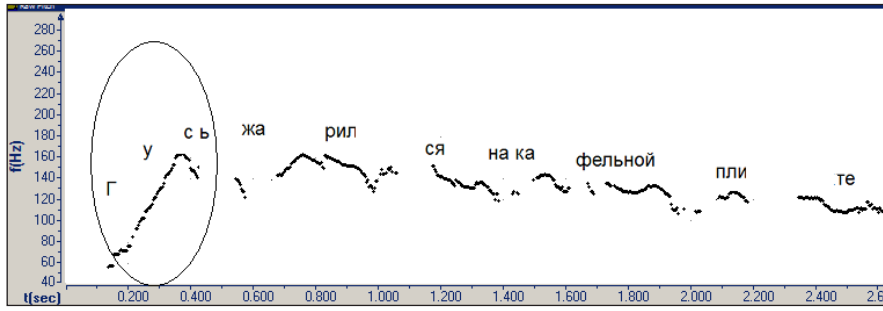


Рис. 8. Тонограмма примера (5)

Именная группа *гусь* в исполнении В. Самойлова несет подъем тона, за которым следует относительно ровное развертывание частоты вплоть до конца предложения. Начальные гласные существенно растянуты. Описание действий героев просодически противопоставлено описанию домашней обстановки. Это интонационная конструкция ИК-6 ([Брызгунова, 1982, 103–118]) плюс существенное

растяжение артикуляции. На тонограмме словоформа-акцентоноситель выделена овалом. Звучание приобретает «мечтательно-воспоминательный» характер.

Итак, начальная группа-новое в чтении в зависимости от лексической семантики и прагматического контекста может входить в следующие коммуникативные структуры:

- с двойной ремой;
- с простой начальной ремой;
- с начальной эмфатической ремой (простой и двойной);
- структурой Тема—Рема (с простой и эмфатической темой);
- с «фоновой» ремой.

Все структуры отличаются от базовой структуры Тема-Данное—Рема-Новое и вносят в семантику предложения и текста дополнительные смыслы.

3. Двойная рема в речи дикторов радио и телевидения

Кроме актерского чтения художественных текстов, специфический интонационный контур с двойной ремой широко используется в речи дикторов радио и телевидения при сообщении новостей. Этот риторический прием мы наблюдаем с семидесятых годов прошлого века. Очевидно, он использовался и до начала наблюдений. Специфика чтения определяется структурой текста, который читает диктор: сегмент, обозначающий новое, помещен автором текста в начало предложения, и в этом же предложении введенному в рассмотрение объекту или событию, дается определенная характеристика. Особенностью речи дикторов служит повышенная по сравнению с предложениями, начинающимися с темы, начальная частота основного тона. Это объясняется тем, что в предвидении близкого падения диктор «набирает высоту», которая обеспечивает достаточно рельефный перепад частот на ударном слоге ацентоносителя первой ремы, расположенного близко к началу предложения. Акцентоноситель второй — конечной — ремы несет второй нисходящий акцент. Рассмотрим один пример из программы новостей:

(6) *Сократить сбор разведданных обещает Барак Обама.*

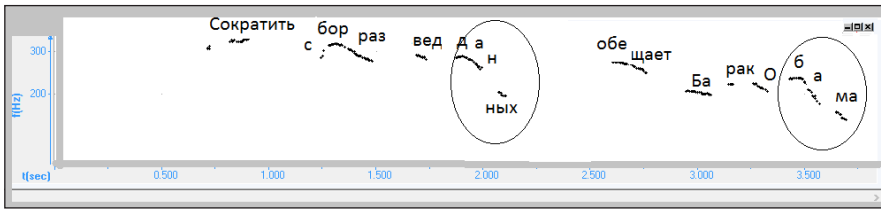


Рис. 9. Тонограмма примера (6).

Тонограмма демонстрирует падение на ударном слоге акцентоносителя начальной ремы словоформы *данных* и второе падение на акцентоносителе второй

ремы *Обама*. Слова-акцентоносители выделены овалами. Перед нами типичная структура с двойной ремой со специфически «концентрированным способом» подачи информации: начальная рема вводит в рассмотрение новый объект — сокращение сбора разведданных, вторая рема содержит сообщение о том, что пообещал сократить разведывательную деятельность не кто иной, как президент Обама.

***

Анализ интонационной структуры массива предложений с начальной группой-новым подтверждает гипотезу И. И. Ковтуновой о совмещении в таких предложениях двух сообщений: первого, вводящего в рассмотрение новый объект или событие, и второго, дающего в том же предложении этому объекту определенную характеристику. Между тем в анализ соответствующих этой семантике коммуникативных структур внесены некоторые уточнения. Предложения с начальной группой, обозначающей новое, реализуется в речи чтецов и дикторов в зависимости от прагматического контекста и лексической семантики предложения в виде следующих коммуникативных структур: двойной ремы; одиночной начальной ремы, простой и эмфатической; начальной темы, простой и эмфатической, и «фоновой» ремы. Структурой, которая наиболее эффектно выражает идею о совмещении в одном предложении двух сообщений, становится структура с двойной ремой. Эта структура весьма частотна в речи актеров, дикторов и лекторов, однако в неподготовленной речи она не используется. В спонтанной речи заготовленное пропозициональное содержание, разбивается на два речевых акта. Таким образом, на примере анализа коммуникативных структур особого типа предложений было показано, что обращение к реальным образцам чтения, широкий охват материала и использование современных средств инструментального анализа позволяют увидеть большее разнообразие комбинаций тем и рем, реализующих информационную структуру с новым в начале, чем это виделось ранее с использованием интроспекции.

Литература

1. *Брызгунова Е. А.* (1982) Интонация, Русская грамматика, том 1, Наука, Москва, сс. 103–118.
2. *Ковтунова И. И.* (1976) Современный русский язык. Порядок слов и актуальное членение предложения, Просвещение, Москва, сс. 103–118.
3. *Ковтунова И. И.* (1979) Структура художественного текста и новая информация, Наука, Москва, сс. 262–274.
4. *Янко Т. Е.* (2008a) Просодические средства эмфазы. Языки славянских культур, Москва, сс. 658–668.
5. *Янко Т. Е.* (2008b) Интнационные стратегии русской речи в сопоставительном аспекте Языки славянских культур, Москва.
6. *Янко Т. Е.* (2012) К классификации падежных и предложно-падежных конструкций. Критика и семиотика, Вып. 17, сс. 155–164.
7. *Chafe W.* (1976) Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View, in Subject and Topic, Academic Press, New York, pp. 25–55.

References

1. *Bryzgunova E. A.* (1982) Intonation [Intonatsija], Russian Grammar [Russkaja grammatika]. Vol. 1, Nauka, Moscow, pp. 103–118.
2. *Chafe W.* (1976) Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View, Subject and Topic, Academic Press, New York, pp. 25–55.
3. *Kovtunova I. I.* (1976) Modern Russian. Word order and theme-rheme division of a sentence [Sovremennyj russkij jazyk. Porjadok slov i aktual'noe chlenenije predlozhenija]. Prosveshchenije, Moscow, pp. 103–118.
4. *Kovtunova I. I.* (1979) Fiction text structure and the new information [Struktura hudozhestvennogo teksta i novaja informatsija], Text syntax [Sintaksis teksta], Nauka, Moscow, pp. 262–274.
5. *Yanko T. E.* (2008a) Emphasis prosody [Prosodicheskie sredstva emfazy], Phonetics and non-phonetics [Fonetika i nefonetika], Jazyki slavjanskih kul'tur, Moscow, pp. 658–668.
6. *Yanko T.* (2008b) Intonatsionnye strategii russkoj rechi v soposnaviteljnom aspekte [Intonational strategies of the Russian speech from a contrastive perspective]. Jazyki slavjanskih kul'tur, Moscow.
7. *Yanko T. E.* (2012) Taxonomy of prepositional and case constructions: the prosodic data [K probleme klassifikatsii padezhnyh i padezhno-predlozhnyh konstruktsij: dannyje prosodii]. Criticism and semiotics [Kritika i semiotika], Vol. 17, pp. 155–164.

ВИД РУССКОГО ГЛАГОЛА И МАШИННЫЙ ПЕРЕВОД

Цангенфайнд Р. И. (r.zangenfeind@lmu.de)

Зонненхаузер Б. (basonne@lmu.de)

Мюнхенский университет, Мюнхен, Германия

Ключевые слова: вид глагола, машинный перевод, неоднозначность, семантический/синтаксический признак, русский язык, английский язык, немецкий язык, турецкий язык

RUSSIAN VERBAL ASPECT AND MACHINE TRANSLATION

Zangenfeind R. (r.zangenfeind@lmu.de)

Sonnenhauser B. (basonne@lmu.de)

University of Munich, Munich, Germany

Rule-based machine translation still offers some very beneficial facets for linguistic theory, because by implementing rules on the computer linguistic theory can be verified in practice. One of the most intricate problems for machine translation is grammatical aspect in Russian when it has to be translated into a language either lacking aspect or having a different aspect system. On the categorical level, aspect has only approximate equivalents in non-Slavic languages, such as the progressive form in English, for instance. In addition, language-internally, its semantics and interpretation cannot be sufficiently captured with only one specific characteristic feature. In this paper, we aim at establishing a basis for the machine translation of the Russian aspect. To do so, we discuss an approach to describe the interaction of verb and aspect semantics in a systematic way. Moreover, we describe a possible annotation for the aspectual information that is provided by further lexical components contributing to the meaning computation. This allows for the formulation of rules for machine translation into target languages where the grammatical category of aspect is realized differently or not present at all.

Keywords: grammatical aspect, machine translation, ambiguity, semantic/syntactic features, Russian, English, German, Turkish

0. Introduction

While statistical machine translation has made great progress over the last years, rule-based machine translation still offers some very beneficial facets. The great virtue of formulating and implementing rules for machine translation instead of using a pure statistical approach is that a rule-based approach is a precious source for theoretical linguistics, cf. Iomdin (2003, 2008), and Apresjan et al. (1989:285).

If dictionaries and rules are implemented in an appropriate way, the computer will be able to produce correct translations. If it does not, it is obvious that the dictionaries or rules have to be improved or new rules have to be added to the system. Thus, the knowledge of rules that describe natural language will be widened and the theory of linguistics will be augmented. This means that even "negative linguistic material" in the form of incorrect translations of a rule-based machine translating system will help to improve linguistic theory.

Apresjan et al. (1989:285) point out that the computer makes mistakes of a different kind from those that a human translator makes. Thus, unique negative linguistic material is provided. Iomdin (2003) gives an example how erroneous automatic parsing of a Russian sentence leads to a wrong translation into English. The examination of this sentence and its syntactic structure reveals a special syntactic property of a group of Russian nouns (*ideja* 'idea' etc.), concerning copulative sentences, that another group of nouns (*tsel'* 'purpose' etc.) doesn't have. By introducing a specific syntactic feature for the according lexemes the parser can be fixed and the sentence is translated correctly.

An especially difficult problem for machine translation is the analysis of the various meanings of Russian verbal aspects. This is a field where rule-based machine translation can be very helpful if appropriate rules are formulated, implemented and verified at the computer. In this paper, we want to discuss the problems of language-internal aspect interpretation and present steps towards rules for machine translation of aspect, especially from Russian to English.

1. Rules for Aspect?

Since aspect interpretation is context-driven and to a large degree subject to pragmatic reasoning, a statistical approach runs into troubles from the very beginning. Gaining statistically valid results for all the possible interpretations would require an immensely large parallel corpus. This makes a rule-based approach look more promising. However, formulating rules for the interpretation and translation of Russian aspect is a rather intricate problem for at least two reasons: this is a highly polysemous category, as can be seen from the numerous readings and sub-readings listed in grammars and textbooks for both aspects, and it has hardly one-to-one correspondences in other aspect languages.

1.1. Interpretation and Correspondences

The multiple interpretations for the imperfective (ipf) aspect can be classified, among others, as 'actual-processual', 'conative', 'habitual', 'atemporal', 'general-factual' and 'durative'. Some readings for the perfective (pf) aspect are the 'event' reading, the 'perfect' and the 'pluperfect' reading. These readings are largely influenced by context. But even considering context, it is not always clear, which interpretation to choose, i.e. which interpretation might be the 'right' one. This makes it quite hard to formulate a common semantic basis for the pf and ipf aspect.

Grammatical aspect is present in other languages as well, e.g. in English and Turkish: At first sight, English *-ing* and Turkish *-iyordu* (*-iyor*=progressive, *du*=past) seem to correspond to the ipf aspect, which would leave the English simple form and the Turkish unmarked past (*-di*) as equivalents to the pf aspect. Such correspondences would simplify the problem of machine translation a lot. But while English uses the progressive form for the actual-processual reading, there is no one-to-one correspondence in the other cases. The habitual interpretation is rendered by the simple form, cf. (1), as is the durative reading, cf. (2). The same holds for the atemporal and the general-factual interpretation, while the conative reading can only be expressed by lexical means.

(1) *At night he **played** with guitarist Luther Perkins and bassist Marshal Grant.*
(http://en.wikipedia.org/wiki/Johnny_Cash, 9.1.2014)

(2) *From 1969 to 1971, Cash **starred** in his own television show […]*
(http://en.wikipedia.org/wiki/Johnny_Cash, 9.1.2014)

Pretty much the same holds for Turkish: *-iyordu* is used for the actual-processual interpretations, the unmarked past for the durative and general-factual interpretations. In addition, Turkish has one further aspect marker, which is used for atemporal and habitual readings, the so-called 'aorist', cf. (3):

(3) *Daha 4 sene öncesine kadar Play Station'da sırf Gerrard'ı kontrol etmek için Liverpool'u **seçerdim**, şimdi beraber oynuyorum.*
(Luis Suarez; http://fotogaleri.hurriyet.com.tr, 10.1.2014) 'Until four years ago I chose Liverpool on the Play Station, just to have Gerrard under control, now we play together.'

As regards the Russian pf aspect, it is expressed in English and Turkish mainly in terms of tense.

Thus, even though English and Turkish have a morphological category of aspect, there is no one-to-one correspondence to Russian. Comparing the semantic range of the Russian, English and Turkish aspect markers, we get the relations illustrated in table 1. German, which does not have a morphological category of aspect, has to rely on lexical and syntactic means:

Table 1. Relations of aspect markers in different languages

Russian	Turkish	English	German
pf	*-di*	simple form	Ø
ipf	*-ir(di)*		
	iyor(du)	*-ing*	

In order to be able to eventually formulate rules in an 'if-then'-format, thus, the following two main problems have to be solved: (i) specify the 'if'-part by language-internally figuring out the relevant interpretation, and (ii) specify the 'then'-part by cross-linguistically figuring out the corresponding equivalent expression. The prerequisite for both is a well-formulated semantic description of aspect.

2. Aspect Semantics

Since it is not possible for to rely on formal equivalences, translation has to take into account the content side. What machine translation cannot achieve is the transfer of specific interpretations since these take into account also extra-linguistic knowledge. What machine translation can achieve, is the transfer of semantically coded meanings. This amounts to the difference between polysemy as the availability of various interpretations for one form and ambiguity as the existence of clearly distinct meanings for one and the same formal expression. This is well-known also from lexical semantics[1]. What is needed in a first step is, thus, a semantic analysis of aspect that is able to distinguish between ambiguity and polysemy.

2.1. Polysemy and Ambiguity

One possible way of systematizing aspect interpretations in terms of ambiguity and polysemy is provided by the analysis developed in Sonnenhauser (2004, 2006), based on the combination of a selection-theoretic (Bickel 1996) and time-relational (Klein 1995) account. According to this analysis, aspect operators select, and thereby assert, specific part(s) of the event structure encoded by the verb. Assuming a tripartite event-structure (Moens, Steedman 1988), verbs may encode (i) dynamic phases 'φ_{dyn}' (preparatory processes), (ii) boundaries 't' (culmination points) and (iii) static phases 'φ_{stat}' (consequent states), depending on the eventuality they refer to. By selecting and asserting some part of the coded event structure, aspect establishes a relation between the topic time interval I(TT) (the time the assertion is about) and the event time interval I(e) (that part of the run time of the denoted event that is selected by the aspect operator).

---

[1] Cf. the German form *Bank* which has at least three meanings: 'bank', 'bench' and 'river bank'. Each of these meanings has its own range of interpretations, i.e. 'bank' may be interpreted as the financial institution, the building, the system, and the like. When it comes to translation, it is not these specific interpretations that are crucial but the three distinct meanings.

The pf aspect can be described by the fact that the boundaries of the event-structure are included in the topic time (a more detailed account is provided in Sonnenhauser 2006, 2009). These boundaries are specified in the course of interpretation: the interval may be closed to both sides, i.e. the initial and final points are part of the interval, it may be open to the right or open to the left, i.e. the initial point is part of the interval whereas the final point is excluded and vice versa. This is illustrated with the example in (4a), which can be interpreted in three ways and thus be translated into English as in (4b–d):

(4)	a.	*Ja emu **dala** knigu.*	
	b.	*I **gave** him the book [and then …]*	I(TT) closed
	c.	*I **have given** him the book [and now …]*	I(TT) open to the right
	d.	*[After] I **had given** him the book*	I(TT) open to the left

For the ipf aspect the following relations between topic time interval and event time interval are relevant:

(5) a. $I(TT) \subset I(\varphi_{dyn})$
*Kogda on voshel, ona **chitala** knigu.* 'When he came in, she **was reading** a book.'
($I(\varphi_{dyn})$: the time interval of her reading the book, covering only this process excluding beginning or end; I(TT) is included in the reading-process and specified by the moment when he came in)

b. $I(TT) = I(e)$
*Ona **rabotala** v universitete.* 'She **worked** at the university.'
[= She was employed there.]
(I(e): the time interval when she was employed at the university; I(TT) runs exactly parallel to the time interval of her working at the university)

c. $I(TT) \supset I(e)$
*Ona uzhe **rasskazyvala** emu ètu istoriju.* 'She **has** already **told** him this story.'
(I(e): the time interval of her telling the story; I(TT) includes the complete story-telling event)

It is these ambiguities that are decisive for the purposes of machine translation; both the structures underlying the representations and the specific interpretations can be neglected.

2.2. Cross-Linguistic Evidence

The justification for postulating the three specifications for the pf aspect is provided not only on language-internal grounds, but also by the fact that these relations can be morphologically coded in other languages, which render it mainly in terms of temporal distinctions. Table 2 illustrates this for Russian, English and German,

with the brackets indicating the boundedness-characteristics of the intervals. Note that these correlations hold for the past tense.

Table 2. Ambiguity of pf aspect

semantics	interpretation	Russian	English	German
group I_{pf} TT closed: [---τ---]	eventive	pf	simple past	imperfect / perfect
group II_{pf} TT right open: [---τ---[	perfect (existential, current relevance, extended now, etc.)	pf	perfect	perfect
group III_{pf} TT left open:]--- τ---]	pluperfect	pf	pluperfect	pluperfect

Likewise, the cross-linguistic validity of assuming three basic ipf configurations is suggested by two facts: the three configurations may be coded morphologically in other languages in terms of aspect distinctions, and if coded, they give rise to a similar range of interpretations. This is illustrated in table 3, comparing 'imperfective' grammemes in Russian, English and Turkish (for more details cf. Sonnenhauser 2006)[2]. This indicates that even though aspect is grammaticalized in all three languages, they are by no means equivalent as regards the semantic range of the respective grammemes.

Table 3. Ambiguity of ipf aspect

semantics	interpretation	Russian	English	Turkish
group I_{ipf} TT ⊂ φ_{dyn}	processual, conative	ipf	progressive	-iyordu -mekteydi
group II_{ipf} TT = e	habitual, non-actual, potential, permanent, atemporal	ipf	simple form	-irdi
group III_{ipf} TT ⊃ e	general-factive, durative	ipf	simple form	-di

The ambiguity of the Russian aspects and the cross-linguistic validity of the possible disambiguated configurations are crucial for the question of machine translation in that this provides the basis for stating clearly formulated rules.

2.3. Disambiguation

Disambiguation is achieved by specifying I(TT) in terms of its boundedness-features and—for the ipf aspect—by specifying the relevant part of the Aktionsart

[2] The comparison in table 3 is confined to the past, since group III_{ipf} is not possible for the other tenses. Accordingly, the Turkish forms are specified with the past tense morpheme *-di*.

that is selected and related to this interval. In Russian, this specification is possible mainly by lexical and syntactic means: as regards the ipf aspect, adverbs like *medlenno* 'slowly' or *postepenno* 'gradually' specify I(TT) as open-bounded, adverbs like *ran'she* 'formerly' as unbounded, particles like *uzhe* 'already' as closed-bounded, and hence the interpretation as belonging to group I_{ipf}, II_{ipf}, or III_{ipf} respectively. Concerning the pf aspect, conjunctions like *i* 'and [then]' disambiguate eventive (group I_{pf}) from perfect (group II_{pf}) interpretations, cf. (6a) vs. (6b), adverbials specifying a point in time suggest the pluperfect interpretation (group III_{pf}), cf. (6c), etc.:

(6) a. *Ja **otkryl** mashinu i sel*. (NKRJa) 'I **opened** the car and [then] got in.'
[---τ---]

b. *Zato synok eë v gorode magazin **otkryl**. Vot i radujtes'…* (NKRJa)
'Instead, her son **has opened** a shop in the city. So be glad…'
[---τ---[

c. *On uzhe **otkryl** rot, no tut v komnatu shirokim shagom voshel djadja Kolja*.
(NKRJa) 'He already **had opened** the mouth, but there uncle Kolja entered the room with big steps.'
]---τ---]

As can be seen from tables 2 and 3, for machine translation from Russian to English, German or Turkish it is enough to solve these basic ambiguities. What is rendered by means of the perfect in English or German has the same interpretational range as the 'perfect' / group II_{pf} specification of the Russian pf aspect, what is rendered by means of the *-irdi* suffix in Turkish may give rise to the same variety of interpretations as group II_{ipf} of the Russian ipf aspect. The same reasoning applies to the other ambiguities.

For an automatic disambiguation, the relevant lexical and syntactic means have to be annotated in the lexical entries of lexemes as regards the aspectual information they contribute to the meaning computation. The computation may then proceed in the form of 'if-then' statements along the lines proposed by Vazov (1999), which is also used by Mel'chuk, Wanner (2008) for aspect-establishing rules in the process of German-Russian translation.

3. Towards Rules for Aspect

The machine translation system ÈTAP-3[3] makes use of a system of semantic and syntactic features (e.g. 'DLIT' to characterize a period of time) which provide a lot of information for lexemes that can be useful for the interpretation of aspect.

For our purpose this system of features could be enriched with a part of the classification of predicates by Apresjan (2006). This classification includes 17 classes. Some of them

[3] ÈTAP-3 is a rule-based MT system for translations from Russian to English and vice versa, and also includes some further NLP applications (cf. Apresjan et al. 2003).

exclude certain disambiguation possibilities and/or make others highly probable. For 'dejatel'nosti' ('activities')[4], such as *torgovat'* 'to trade', for instance, the actual-processual and the general-factual readings are ruled out, whereas a durative interpretation is most likely. Other classes, such as 'dejstvija' ('actions'), are a lot less explicit and allow for all possible interpretations. For their disambiguation, further information provided by other aspectually relevant components in the regarded sentence must be taken into account.

Adverbials, particles and conjunctions provide this information[5]. These parts of speech have to be assigned with additional semantic and syntactic features respectively in their lexical entries. Another crucial bit of information is provided by tense. Present tense, for instance, excludes ipf interpretations out of group III_{ipf} and all pf interpretations except for the future interpretation. The combination of all this kind of information can be the basis for the "calculation" of a temporal and aspectual interpretation of the whole sentence.

An example to illustrate which information in a sentence is relevant is given in (7):

(7) *Ran'she ja po vecheram prodelyval èti gimnasticheskie uprazhnenja po pjat' raz*[6].
lit. 'formerly I in evenings do.PAST.ipf these gymnastic exercises each five times'

Most lexemes and phrases in this sentence are important for our interpretation. For all of them the dictionary entries of ÈTAP already provide some important information, which, for our purposes, should be enriched by the following:

- *ran'she* 'formerly' is temporally and referentially (as concerns reference to event) indefinite and thus excludes group I_{ipf} interpretations; appropriate semantic features could be 'temporally indefinite' and 'referentially indefinite'[7]
- *po [vecheram]* 'in [the evenings]': the preposition in this expression—governing a temporal lexeme in the dative case, i.e *po16*[8]—expresses regularity. An adverbial phrase like *po vecheram* 'in the evenings' can be annotated by labeling the preposition *po16* with the feature 'regularity'; thus, it excludes group I_{ipf} and group III_{ipf} interpretations
- *prodelyvat'* '[to] do' is used as a support verb; i.e. it has no semantics, only its aspectual information (=ipf) is relevant

---

[4] The English terms for classes of predicates are taken from Apresjan (2005).

[5] These components correspond to the contextual clues (imperfective and perfective triggers) of Mel'chuk, Wanner (2008).

[6] Example from Bendixen et al. (2005–2012).

[7] The semantic feature 'temporally indefinite' indicates that there is just a vague temporal specification in terms of localization on the time axis. The lists of adverbs with this and other semantic features still must be thoroughly examined; the need for a list of such triggers is pointed out also by Mel'chuk, Wanner (2008:141). 'Referentially indefinite' concerns the selection and assertion of a specific part of the event structure carried out by aspect (cf. section 2.1): adverbs like *ran'she* indicate that there is no specific part of the event structure selected by aspect (some more examples of such features are given in Sonnenhauser, Zangenfeind 2013).

[8] cf. Slovar' russkogo jazyka (1983).

- *uprazhnenie* 'exercise' is the semantic predicate in the sentence and can be labeled as 'zanjatie' ('occupation') according to Apresjan (2006: 83, 86f.); in combination with an ipf support verb such as *prodelyvat'* it allows for group I_{ipf}, II_{ipf} and III_{ipf} interpretations
- *po [pjat' raz]* '[five times] each': the preposition here—governing a noun that can have a numeral as syntactic dependent, i.e. *po20*[9]—expresses distributivity of the verbal complement and allows for group I_{ipf}, II_{ipf}, III_{ipf} interpretations; the preposition *po20* can be labeled with the feature 'distributive'.

Based on this information, the aspectual information given in (7) can be calculated and, thus, disambiguated as follows:

(8) for language-internal disambiguation:
IF predicate has feature 'occupation'
AND IF aspect = ipf
AND IF tense = past
AND IF there is an adverb of 'group II_{ipf}'
THEN 'group II_{ipf}' interpretation

(9) for translation into English:
IF 'group II_{ipf}' interpretation
THEN 'simple form' in English[10]

Formal descriptions like these can be the basis for an implementation in a machine translation system like ÈTAP.[11]

4. Conclusion

In machine translation a rule-based approach for the interpretation and translation of the Russian verbal aspect looks promising when using the combination of a selection-theoretic and time-relational account to systematize the semantics of aspect and its interpretations. This systematization comprises several groups specifying the relation between topic time interval and event time interval. Disambiguation of the semantics of aspect is made possible by annotating all relevant lexemes with specific, aspectually

---

[9] cf. Slovar' russkogo jazyka (1983).

[10] The most adequate translation would be with the habitual construction 'used to'. This specification can be solved by means of language-internal paraphrasing rules and is not necessarily an immediate concern of translation.

[11] Since ÈTAP includes a highly developed Russian-to-English MT system, we intend to implement rules for aspect translation into English in a first step. But beginnings for the implementation of Russian-German translation in ÈTAP have already been made and are developed further by R. Zangenfeind and others. So, in the long run the translation of aspect from Russian to German is also planned.

relevant information. This is the starting point for a possible computational implementation of aspect interpretation. Enriching the system of semantic and syntactic features of the machine translation system ÈTAP with Apresjan's classification of predicates and with additional, more detailed syntactic/semantic features, we discussed the problems of a "calculation" of aspect interpretation and presented steps towards a possible solution.

Our future work will be to develop the necessary system of semantic features for verbs and predicative nouns, adverbials, particles and conjunctions. It is our aim to implement rules for aspect translation in the machine translation system ÈTAP. Besides the practical utility, an implementation in a rule based system has the great virtue to verify the linguistic theory in practice and, with that, to enable an improvement of the theory.

References

1. *Apresjan Ju. D.* (2005), Prolegomena to systematic lexicography, in: Apresjan Ju. D., Iomdin L. L. (eds.), East West encounter: second international conference on Meaning ↔ Text Theory, Moscow, pp. 20–29.
2. *Apresjan Ju. D.* (2006), Fundamental classification of predicates [Fundamental'naja klassifikatsija predikatov], in: Apresjan, Ju. D. (ed.), A linguistic picture of the world and systematic lexicography [Jazykovaja kartina mira i sistemnaja leksikografija], Moscow, pp. 75–110.
3. *Apresjan Ju. D.* et al. (1989), Linguistic Software for ÈTAP-2 System [Lingvisticheskoe obespechenie sistemi ÈTAP-2], Moscow.
4. *Apresjan Ju. D.* et al. (2003), ÈTAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the Meaning ↔ Text Theory, in: Conference Proceedings of MTT 2003, Paris, pp. 279–288, available at http://proling.iitp.ru/publications/.
5. *Bendixen B.* et al. (2005–2012), Russian up to date [Russisch aktuell], Wiesbaden.
6. *Bickel B.* (1996), Aspect, mood and time in Belhare, Zürich.
7. *Dictionary* of the Russian Language (1983), vol. III [Slovar' russkogo jazyka, Tom III], Moscow.
8. *Iomdin L.* (2003), Purpose and Idea: a Lesson Drawn from Machine Translation, in: Conference Proceedings of MTT 2003, Paris, pp. 269–278.
9. *Iomdin L.* (2008), A few Lessons Learned from rule-based Machine Translation, in: Gross G., Schulz K. U. (eds.), Linguistics, Computer Science and Language Processing. Festschrift for Franz Guenthner on the Occasion of his 60th Birthday. London, pp. 171–187.
10. *Klein W.* (1995), A time-relational analysis of Russian aspect, in: Language 71(4), pp. 669–695.
11. *Mel'chuk I. A., Wanner L.* (2008), Morphological Mismatches in Machine Translation, in: Machine Translation, 22, pp. 101–152.
12. *Moens M., Steedman M.* (1988), Temporal ontology and temporal reference, in: Computational Linguistics 14(2), pp. 15–28.
13. *Sonnenhauser B.* (2004), Underspecification of 'meaning': the case of Russian imperfective aspect, in: Proceedings of the ACL-04 workshop on 'Text Meaning and Interpretation'. Barcelona, pp. 89–96.

14. *Sonnenhauser B.* (2006), Yet there's method in it. Semantics, pragmatics, and the interpretation of the Russian imperfective aspect, Munich.
15. *Sonnenhauser B.* (2009), Definiteness and specificity of verbal referents, in: Birzer S., Finkelstein M., Mendoza I. (eds.), Proceedings of the second international Perspectives on Slavistics conference (Regensburg 2006), Munich, p. 115–126.
16. *Sonnenhauser B., Zangenfeind R.* (2013), Towards machine translation of Russian aspect, in: Apresjan V., Iomdin B., Ageeva E. (eds.), Proceedings of the 6th International Conference on Meaning-Text Theory. Prague, pp. 192–201, available at http://meaningtext.net/mtt2013/proceedings_MTT13.pdf.
17. *Vazov N.* (1999), Context-scanning strategy in temporal reasoning, in: Modeling and using context. Lecture notes in computer science 1688, pp. 389–402.

SENTENTIONAL ARGUMENTS AND EVENT STRUCTURE[1]

Zimmerling A. V. (fagraey64@hotmail.com)

Sholokhov Moscow State University for the Humanities,
Moscow, Russia;
Institute of Linguistics of the Russian Academy of Sciences,
Moscow, Russia

This paper is addressed the interaction of subject marking and event structure in languages, which allow sentential arguments in the subject position. In Russian and other Slavic languages sentential subjects share a number of formal and semantic properties with zero subjects with role-and-reference features and with so called oblique subjects, i.e. subject-like arguments marked with an oblique case. I argue that sentential subjects represented by bare that-clauses (Rus. *čto*-clauses) cannot have the roles of Agent/Causer, while zero subjects can. I also argue that the capacity of taking *to, čto P*-clauses, i.e. that-clauses headed by a correlative pronoun *to* serves as diagnostics for a number of verbal classes. Causative predicates like *vynudit', zastavit', sklonitj k čemu-l.* only take *to, čto P*- clauses, but not bare *čto P*-clauses as surface subjects. Factive predicates like *znat', razdražat'* etc. take *to, čto P*-clauses, but not bare *čto P*-clauses as surface subjects while non-factive predicates like *dumat'*, *mereščit'sa* only take bare *čto P*-clauses. Nominal predicatives forming Dative-Predicative-Structures (DPS) with an oblique subject marked with dative case and specified as {+ animate; + referential} split into two groups. Russian DPS predicatives from the *stydno, dosadno, protivno, vse ravno* group only take bare *čto P*-clauses and invariably behave as non-factive verbs in all contexts with an overt oblique subject. Russian DPS predicatives from the *izvestno, neizvestno, stranno, bezrazlično* group both take bare *čto P*-clauses and headed *to, čto P*-clauses, i.e. can be used in factive contexts as well. That means that their sentential argument can both get the status of a fact i.e. verified proposition P, logical truth, and an intentional situation, e.g. subjective evaluation of P, inner vision of P etc. Russian has two expletive elements—*eto* and *to,* but their syntax is different. *Eto* behaves as surface subject of the matrix clause and alternates with oblique subjects and sentential arguments in the subject position while *to* cannot be separated from the complement clause and reaches the subject position only in combination with the CP.

Key words: grammatical subject, sentential subjects, expletive subjects, zero categories, case marking, semantic roles, factivity, facts, situations, event structure

---

[1] The paper is written with financial support from the Russian Ministry of Education and Science, federal project 2685 'Parametric description of grammar systems'. I am grateful to the anonymous reviewers of the 'Dialogue 2014' conference and to Ekaterina Lyutikova and Oleg Belyaev for the valuable comments and criticism. The responsibility for all shortcomings is on the author.

Что сентенциальные подлежащие существуют, всем известно.

0. Introduction

This paper is addressed syntax and semantics of sententional subjects in a number of European languages with the nominative-accusative sentence pattern. I argue that sententional subject arguments are typical for sentences without an Agent NP/DP. Sententional subjects can express semantic roles as Causer or Stimulus with predicates subcategorizing for an animate Patient or Experiencer argument marked with accusative or oblique case. Such predicates denote uncontrolled events, which do not imply any human Agent or volitional Causer. The same languages, notably, Modern Icelandic, Ukrainian and Russian, also make use of zero subjects with similar role-and-reference properties. Both sententional subjects and zero subjects can have the value of External Force. I argue that Ukrainian and Russian sententional subjects with a value of External Force project event structure without a human Causer while some types of zero subjects in these languages project event structure with a human Causer. This contrast is partly due to different parameter settings and partly due to lexical semantics, since sententional and zero subjects are licensed by different groups of predicates.

The paper is organized as follows: in section 1, I render the notion of sententional subjects, in section 2 I discuss subjecthood tests for several Germanic and Slavic languages. In section 3, I analyze event structure projected by predicates licensing sententional subjects and by predicates licensing zero subjects. In section 4, I offer syntactic diagnostics for Russian predicates with a propositional argument.

1. The Hypothesis of Sententional Subjects

The notion of sententional subject is deeply rooted in the European linguistic tradition and, with some stipulations, in the Russian linguistic tradition as well. Many descriptive grammars from the Young Grammarian time took for granted that if a language has grammatical subjects in the nominative case and a nominative NP/DP or a pro-form is lacking, then the subject position must be filled by other expressions, e.g. by an infinite phrase, subordinate clause or a dummy. For instance, Nygaard (1906: 220) promptly says in his Old Norse syntax that 'infinitive is used as subject' in examples like OIce. **hørmulikt** er [$_{IP}$ slíkt <u>at vita</u>] 'it is **sad** to know this', lit. '**sad** is [$_{IP}$ such <u>to know</u>]'. The same is later said about 'subordinate clauses in the subject position' [Nygaard 1906: 252]. Nygaard does not prove either claim, since he believes that if a nominative argument is absent, the subject position must be filled by some placeholder—an idea which was later capitalized in the classical version of the Raising Theory [Perlmutter, Postal 1983][2]. A related approach to Burzio's Generalization

[2] In fact, many Nygaard's examples are questionable, since with Old Icelandic verbs like *þykkja* 'to seem' sententional arguments do not raise to the subject position, as shown in [Zimmerling 2002: 637] while this language licenses nominative objects with an impersonal verb [ibid., 770].

and nominative case marking in the Minimalist Framework has been aptly dubbed 'nominative-first-syntax' by Lavine (2014)[3]. Peškovskij (1938) mentions sentences like Rus. [$_{IP}$ Vozbuzhdat' lyubopytstvo] sil'no l'stilo ego samolybiu '[$_{IP}$ to provoke curiosity] flattered his self-esteem strongly', where the infinitive 'fills in the place of subject' (1938: 203). He calls such sentences 'very rare'. On the contrary, an academic grammar of Russian [Švedova et alii 1982: 94] openly declares that grammatical subject is an obligatory component that is regularly expressed either by a noun in the nominative case, or with 'an infinitive having the value of semantic subject'[4]. Neither Peškovskj nor Švedova acknowledge finite clauses as subjects, except for the case where such clauses serve as titles of quotations [Peškovskij 1938: 202; Švedova et alii 1982: 122]. Though the latter source mentions *in passim* that subordinate clauses occur with putative and affective verbs and lists examples like Rus. Emu mereščilos', [$_{CP}$ čto travit on lisu] 'he fancied/dreamed that he was hunting a fox' [Švedova et alii 1982: 494], no attempt to analyze their status is made—probably because Švedova et alii treat the absence of a nominative subject with verbs like *mereščit'sa* 'to fancy' / 'to dream' as an idiosyncratic, i.e. lexical feature of impersonal verbs. There are, however, numerous counterexamples, where predicates licensing a sententional argument are not specified as impersonal in the lexicon and a nominative argument is lacking, cf. structures with a predicative *nexorošo* (1) and structures with affective transitive verbs like *razdražat'* 'to annoy' (2a), which also take nominative subjects (2b–c).

(1) Rus. [$_{CP}$ čto deti ostalisj golodnye]—nexorošo$_{Pred}$.
'It is bad [$_{CP}$ that children remained hungry].'

(2) Rus. a. Vas'u$_{Acc}$ razdražaet$_{3Sg}$ [$_{CP}$ čto Katia postojanno opazdyvaet].
'it gets on Vasja's nerves [$_{CP}$ that Kate always comes late].'
Lit. 'to-Vasja annoys [$_{CP}$ that Kate constantly comes late].'
b. [$_{NP}$ Postojannye$_{Nom.Pl}$ Katiny opazdanija$_{Nom.Pl}$] razdražajut$_{3Pl}$ Vas'u$_{Acc}$.
'[$_{NP}$ Kate's constant late arrivals$_{Nom.Pl}$] annoy$_{3Pl}$ Vasju$_{Acc}$.'
c. Katia$_{NomSg}$ /Devuška$_{NomSg}$ razdražaet$_{3Sg}$ Vas'u$_{Acc}$ svoimi$_{Instr.Pl.}$ opazdanijami$_{Instr.Pl.}$.
'Katia$_{NomSg}$ /The girl$_{NomSg}$ annoys$_{3Sg}$ Vasia$_{Acc}$ by her late arrivals$_{Instr.Pl}$.'

Testelets (2001: 318) identifies sententional arguments in structures like (1), (2a) as subjects or expressions behaving as grammatical subjects, which seems to be a consequent solution. A further problem is whether sententional arguments express semantic roles or just behave as placeholders in the subject position. The intuition for (1) and (2a) is different. While CP [$_{CP}$ *čto deti ostalisj golodnye*] filling a valency slot

---

[3] It is not clear whether sententional arguments and other non-standard subject-like expressions filling the subject position get nominative case, but this is merely a framework-internal issue.

[4] The notion of semantic subject in this definition is vague. It is difficult to assign infinite phrases in sentences like Rus. [$_{IP}$ *Zanimat'sa sportom*]—***vredno*** 'It is unhealthy to go in for sports', lit. [$_{IP}$ to go in for sports] is **unhealthy**' any semantic role except for 'ability to conduct the process designated by the verb'.

of a predicative *nexorošo* 'bad that P' can hardly be specified more detailed than 'Situation P', the CP [$_{CP}$ *čto Kat'a postojanno opazdyvaet*] 'that Kate constantly comes late' filling a valency slot of the transitive verb *razdražat'* 'annoy' in (2a) seems to express the same semantic role of Stimulus as the nominalization [$_{NP}$ *Postojannye Katiny opazdanija*] 'Kate's constant late arrivals' in (2b) or standard NPs like *Kate/the girl* in (2c). The second argument in (2a–c) is marked with accusative and likely has the same role (likely—the role of Experiencer) with all subjects[5]. Hence, it seems that if a verb taking a non-sententional argument with a given semantic role also takes a sententional argument, the latter can inherit the same role value. Belletti & Rizzi (1988) argue that sententional arguments are always internal, since they occur with so-called psych verbs, i.e. predicates without an Agent subject; that means that if CPs/IPs take the position of surface subject, they are nevertheless derived subjects raised to subject position in the absence of categories standing higher in subject hierarchy—Agent subjects in the direct case or semantic subjects in an oblique case etc. This issue will be addressed in this paper later.

2. Subjecthood Tests and Expletive Subjects in Germanic Languages

The theory of sententional subjects is supported by the observation that predicates taking sententional arguments also take non-sententional ones, cf. (2a) vs (2b–c) above. Germanic languages add a special issue—some predicates taking sententional subjects can and in certain configurations must take expletive subjects like Eng. *it*, Da., Sv, *det*, Ger. *es*[6]. In English and in Mainland Scandinavian languages expletive subjects occur in structures like (3a), where they anticipate a postponed that-clause, but are absent if that-clauses are preposed, The ill-formedness of structures like Eng. (3b) indicates that preposed CPs take subject position while examples like Eng. (3c) show that anticipatory *it* has subject properties too, since it is preserved with inversion and other syntactic transformations.

(3) Eng. a. *It* is suprising, [$_{CP}$ that John knows about you].
b. [$_{CP}$ that John knows about you] is *(it) suprising.
c. Eng. a. Is *it* suprising, [$_{CP}$ that John knows about you]?

---

[5] Semantic roles associated with Russian *čto*-clauses are analyzed in Kniazev (2012).

[6] Cross-linguistically, expletive elements also occur in personal clauses, where they anticipate sententional arguments. This is possible for Sw., Da., Norw. expletive *det* and Ger. expletive *es*, and not typical of Eng. *it*—except for examples like Eng. The Foreign Secretary made it_i clear [$_{CP}$ that the President is not prepared to make any decision regarding this problem]i. In German linguistics this function of expletive words is called 'Korrelat' (correlate). In this paper, I concentrate on subject uses of expletive elements. If the same lexical element, as Da. *det* or Ger. *es* is used both as subject and as correlate, I analyze subject and non-subject uses as separate syntactic categories.

If the expletive *it* is in subject position in (3a) and (3c), then the CP [$_{CP}$ that John knows about you] is an adjunct in (3a) and (3c), though it apparently behaves as subject in (3b), where the insertion of expletive *it* is impossible. Partee (1979: 17–22) assumes that (3b) has sentential subject, but challenges the idea earlier proposed by O. Jespersen and G. Curme that (3a) has sentential subject too. She argues that if surface subject is defined by substitution (i.e. structure preservation criterion), 'anticipatory *it* will necessarily be treated as subject whether it is considered as a part of the underlying subject or transformationally introduced to its place' (ibid., 21). An alternative approach when surface subject is defined on the basis of person-and-number agreement rule[7] does not work. Jespersen's claim that the sentential argument is invariably selected as surface subject irrespective of the fact whether it is postponed or preposed can be saved only under the assumption that *it* in (3a) is not inserted until postposing of the CP takes place. This is unlikely both on empiric reasons, since the postposition of that-clauses is their normal position in right-branching languages like English, and on theoretical reasons, since there is no evidence that postposition of that-clauses takes place at all. It is easier to analyze preposed that-clauses as fronted, i.e. moved from postposition to the preverbal position, and conclude that insertion (in recent terminology, merger) of expletive elements like *it* is only possible if the CP is <u>not</u> fronted.

Mainland Scandinavian languages display the same complementary distribution of expletive and sentential subjects in structures like (3a–c). The expletive *det* is obligatory in (4a) and (4c) and ruled out in (4b), where the CP is preposed. This prompts that a) *det* and CP alternate in the surface subject position, b) the same predicates take expletive or sentential subjects with different word order and configuration.

(4) Da. a. *Det* er mærkeligt$_{Pred}$ [$_{CP}$ at Jens ikke kender hende].
'It is strange [$_{CP}$ that Jens does not know her].'
b. [$_{CP}$ at Jens ikke kender hende] er *(*det*) mærkeligt$_{Pred}$.
'[$_{CP}$ that Jens does not know her] is strange'.
c. Er *det* ikke mærkeligt$_{Pred}$ [$_{CP}$ at Jens ikke kender hende]?
'Isn't it strange [$_{CP}$ that Jens does not know her]?'

The agreement criterion is less telling in Danish, Swedish or Norwegian, since the impoverished verbal morphology of these languages does not show whether sentential subjects agree with the verb in the 3rd person, which is not marked overtly. The linear position criterion gives mixed results. Unlike English, Danish, Swedish and Norwegian are standard verb-second languages, where the preverbal position (XP) is not reserved for subject NPs while the unique position specific of subject NPs and pro-forms is located <u>after</u> the finite verb[8], but before the general negation; object NPs

---

[7] For languages like English, German, Icelandic or Russian where verbs taking sentential subjects are morphologically marked as standing in 3rd person singular. For Danish, Swedish and Norwegian which have impoverished verbal morphology, the default agreement value is just 3rd person (defined in syntax since morphological person markers are missing as well).

[8] This parameter differs the Mainland Scandinavian type from other verb-second languages, like Kashmiri or German, where subject NPs do not get a canonic position after the verb and are

are placed after infinite verbs. This gives the main clause order XP–V_{fin}–**NP**$_{Sub}$–Neg/AdvSent–V_{inf}–**NP**$_{Obj}$, cf. Zimmerling (2002: 279). Expletive elements like *det* in diagnostic contexts with general negation, cf. (4c), came up in the position specific of subject NPs [Ekerot 1995] while sententional subjects, cf. (4b), are possible only in XP, where both fronted subjects and fronted objects/adjuncts occur. To complete the picture, one must mention that Danish, Swedish and Norwegian, unlike English, but like Russian or Icelandic have a parameter licensing XP-fronting of the nominal predicative in sentences like *(it) is good that P* In this case the expletive does not show up in main clauses, cf. Sw. *Det* är **bra** [$_{CP}$ at han kommer] 'It is good that he comes', but Sw. # **bra**$_i$ är t_i [$_{CP}$ at han kommer], lit. '**good** is [$_{CP}$ that he comes]', cf. [Zimmerling 2002: 738][9].

The substitution criterion gives conclusive proof that expletives like Da. *det* in (4) act as surface subjects, since their position is preserved in embedded structures: (5c) with a preserved expletive is well-formed, while (5d) without the expletive is bad.

(5) Sw.

Basic structure	**Derived structure with embedding**
(5a) *Det* är bra [$_{CP}$ at han kommer] ~ **bra**$_i$ är t_i [$_{CP}$ at han kommer]. 'It is good that he is coming.'	(5c) [$_{CP}$ Om *det* är bra [$_{CP}$ at han kommer]] är jag Karl XII. '[$_{CP}$ If it is good [$_{CP}$ that he is coming]], I am Charles XII.'
(5b) [$_{CP}$ at han kommer] är *(*det*) bra. '[$_{CP}$ That he is coming] is good'	(5d) *[$_{CP}$ Om [$_{CP}$ at han kommer] är bra] är jag Karl XII.

This distribution proves that neither expletive nor sententional subjects can be eliminated from the description of Mainland Scandinavian languages, since there are both structures where the expletives are obligatory—main clauses without fronted CPs, cf. (5a), or IPs[10], structures with embedding (5c), and structures with fronted CPs, cf. (5b), or IPs[11], where expletives are ruled out. In the latter case fronted CPs / IPs act as surface subjects, in the first case they must be analyzed as non-arguments, i.e. adjuncts.

Some theorists have tried to get rid of sententional subjects in Universal Grammar and claimed that the subject position in the process of derivation is actually filled not by IPs/CPs but by some zero categories coindexed with them: these zero categories are allegedly made visible in some languages as overt expletive elements like Eng. *it* in (3) or Da., Sw. *det* in (4)–(5). The idea that expletives have zero counterparts in the same or other languages is not new, but until recently it has not been combined with the denial of sententional subjects. The elimination of sententional subjects and other

placed in the so called middle field, i.e. scrambling area between the finite and infinite verbs: XP – V_{fin} {$_{Scrambling}$S + O + Adv} V_{inf}. Cf. [Bhatt 1999], [Zimmerling 2013a: 188–189] for details.

[9] However, XP-fronting of nominal predicatives gives stylistically marked sentences and is blocked with most predicatives in Danish.

[10] Cf. Da. *Det* er godt [$_{IP}$ at drikke øl] 'It is good [$_{IP}$ to drink beer]' ~ *godt er [$_{IP}$ at drikke øl]

[11] Cf. Da. [$_{IP}$ at drikke øl] er *(*det*) godt. '[$_{IP}$ To drink beer] is good'.

subject-like expressions alternating with one and the same predicate is desirable, but ascribing subject properties to zero categories coindexed with IPs/CPs rather creates problems than solves them. The constraints on merging zero forms into subject positions, as Germanic languages show, are linked with overt expletives acting as surface subjects, not with silent categories allegedly coindexed with IPs/CPs. This contradicts the initial assumption that overt and silent expletives are just two sides of the same category. If, on the contrary, overt expletives in clauses with postponed IPs/CPs and zero forms coindexed with IPs/CPs are categories of a different sort, we are left back with a version of traditional analysis in terms of sententional subjects.

3. Sententional, Expletive and Oblique Subjects in Russian

Descriptive grammars of Russian and most other Slavic languages state that they lack expletive elements[12], so the alternation of expletive vs sententional subjects should not be a problem of Russian syntax. This view has been challenged in [Zimmerling 2009; 2012], where the syntax of Rus. non-referential non-agreeing element *eto* 'it' is discussed. The non-referential non-agreeing prosodically weak *eto* (dubbed 'semi-expletive eto' in [Zimmerling 2009]) freely combines with that-clauses, cf. (6a), but in one special case where the matrix predicate belongs to the class of the so called 'category-of-state forms', i.e. non-agreeing nominal predicatives selecting a dative subject, cf. *mne grustno* 'I am sad', *mne protivno* 'it makes me sick', *mne stranno, udivitel'no* 'It seems strange/suprising to me', *mne jasno, očevidno* 'It is clear/evident (to me)' etc., the combination of semi-expletive eto + CP is blocked in the presence of a dative subject (6b), though neither a combination dative subject + semi-expletive *eto*, cf. (6c) nor a combination dative subject + CP, cf. (6f) are ruled out. If there are no other candidates for the subject position, CP acts as surface subject—both when it is preposed (6d) and postposed (6e).

(6) Rus. a. *Eto* udiviteljno$_{Pred}$, [$_{CP}$ čto pogoda ne isportilas'].
'It is suprising [$_{CP}$ that the weather did not worsen].'
b. *__Mne__ *eto* udiviteljno$_{Pred}$, [$_{CP}$ čto pogoda ne isportilas'].
c. __Mne__ *eto* udiviteljno$_{Pred}$.
d. [$_{CP}$ čto pogoda ne isportilas'], udivitel'no$_{Pred}$.
e. Udivitel'no$_{Pred}$, [$_{CP}$ čto pogoda ne isportilas'].
f. __Mne__ udivitel'no$_{Pred}$, [$_{CP}$ čto pogoda ne isportilas'].

This distribution is straightforwardly explained if all three sorts of expressions in (6)—dative subjects, semi-expletive *eto* and sententional arguments are derived subjects, i.e. internal arguments of Russian predicatives promoted to the vacant subject position according to some hierarchy of arguments. This analysis has been outlined by Zimmerling (2009; 2012) who postulates the following hierarchy for Russian Dative-Predicative-Structures (DPS):

---

[12] Overt expletive subjects are attested in Upper Sorbian, arguably due to German influence, cf. Zimmerling (2002: 541; 750).

(i) Dative subject >> sententional subject >> semi-expletive *eto*.

If (i) holds for Russian DPS, dative DPS subjects with the role of Experiencer have a priority over sententional arguments: the latter are chosen as subjects only if dative subjects are absent. If neither dative nor sentential subjects are present, semi-expletive *eto* is selected as subject. Letuchiy (2014) accepts the hypothesis that dative DPS arguments and sentential arguments alternate in the surface subject position, but argues that sententional arguments have a priority over dative subjects, so the hierarchy according to him is (ii)[13];

(ii) Sententional subject >> dative subject >> *eto*.

In the perspective of this paper, the choice of subject hierarchy (i) vs (ii) is not relevant, but this issue is important for the description of Russian DPS. Russian has ca. 300 non-agreeing nominal predicates capable of forming DPS, all of them select dative subjects specified as {+animate' + referential}. Roughly one third of them (cf. *udivitel'no, izvestno, stranno*, *stydno, žal'*, *protivno*) select that-clauses. If sententional arguments have a priority over dative subjects, one has to prove that dative arguments of DPS predicatives take object positions when CPs are present, cf. (6f). It is unclear whether this can be done, since dative subjects are thematic, regularly fronted elements which do not behave like other arguments of DPS predicatives. The absence of an overt dative argument in the presence of a CP in (6e) and (6d) is satisfactory explained by a shift from a overt referential Experiencer (***Mne** udivitel'no, čto P*—'situation P seems strange to **some referential X**') to a silent non-referential Experiencer (∅ *udivitel'no, čto P*—'situation P will seem strange **to every X**'). Hence, the hierarchy (i) seems to give a more economic description of Russian DPS sentences than the hierarchy (ii).

Apresian (1985: 304) lists sententional complements, *eto* and pronominal correlative *to* $_{\text{3Sg.Nom-Acc.N}}$ 'that' as categories that can fill the subject position of DPS predicatives. However, *eto* and *to* hardly have the same syntax, since expletive *eto* can take distant position, cf. (6a) above, stands before or after CP, while unstressed[14] expletive *to* cannot be separated from the subordinate clause it heads: *mne udivitel'no* ***to***$_{\text{i}}$ [$_{\text{CP}}$ *čto on eš'o ne sdals*'a]$_{\text{i}}$ 'It$_{\text{i}}$ seems strange to me [$_{\text{CP}}$ that he still did not give up]$_{\text{i}}$', ****to***$_{\text{i}}$ *mne udivitel'no* [$_{\text{CP}}$ *čto on eš'o ne sdals'a*]$_{\text{i}}$. Meanwhile, expletive *eto* has a strong propensity for fronting (like Eng. *it* or Da. *det*) which is not relevant for expletive *to*, since the latter always immediately precedes its CP. Finally, *to,* unlike *eto,* does not alternate with sententional arguments and oblique dative subjects. Therefore, it is not part of subject

---

[13] The view that those Russian DPS predicatives which have sententional arguments are not impersonal and take sentential arguments and pronominal elements *eto* and *to* as surface subjects has earlier been defended by Apresian (1985: 304).

[14] A syntactic homonym, stressed *TO,* can be separated from CPs if it has contrastive stress, though such sentences do not look natural: Rus. *?TO mne udivitel'no* [$_{\text{CP}}$ *čto on eš'o ne sdals'a*] 'It seems strange to me [$_{\text{CP}}$ that he still did not give up].' Stressed pronoun *TO*, unlike unstressed to, can be enhanced by the enclitic to2 'emphatic theme', in combination with emphatic proclitic i: *TO=to*$_{2}$ *mne* ***i*** *udivitel'no* [$_{\text{CP}}$ *čto on eš'o ne sdals'a*].

hierarchies like (i)-(ii) and likely not part of the main clause structure—this issue is to be discussed below in section 4.

4. Zero Subjects and Event Structure

Many languages with sententional subjects including Russian, Ukrainian, Modern and Old Icelandic also have zero subjects with the role-and-reference properties. Russian, Ukrainian and Icelandic zero subjects are non-referential Agents/Causers, which can be specified both as {+ animate} and {− animate} in constructions of a different type, cf. Lavine (2014), Zimmerling (2013). Many predicates license both constructions with {+ animate} and {− animate} zero subjects in the same language, cf. (7a–b) and (8a–b). In both cases zero subjects exhibit some kind of agreement with the predicate which has been shown for Russian impersonals by Mel'čuque (1979).

(7) Rus. a. $\varnothing^{3Sg}$ {− animate} **Lodku**$_{AccSg}$ oprokinu-l-o$_{3Sg.N.Pst}$ (vetrom$_{Instr.Sg}$).
'The boat turned over (due to a puff).'
b. $\varnothing^{3Pl}$ {+ animate}**Lodku**$_{AccSg}$ oprokinu-l-i$_{3Pl.Pst}$ (by a puff$_{Adv}$).

(8) Icel. a. $\varnothing^{3Sg}$ {− animate} **Bátunum**$_{DatPl}$ hvolf-d-i$_{3Sg.Pst}$ (*viljandi).
'The boat turned over (*by purpose).'
b. $\varnothing^{3Sg}$ {+ animate} **Bátunum**$_{DatPl}$ var$_{3Sg}$ hvolf-t$_{.Prt.Pst.N.Sg.}$ viljandi.
'The boat has been turned over by purpose <by some people>).'

Zero subjects specified as {− animate} typically occur in transitive impersonals like (7a), (8a) in sentences denoting processes not controlled by any human Agent. Their role can be defined as non-human Agent or as Causer, if non-human Agents are not accepted in semantic description, cf. Lavine (2014). The silent Agent/Causer argument is paired in transitive impersonals with an overt argument having the role of Patient: the case-marking of the latter depends on the verbal government—in the Russian example (7a) the verb *oprokinut'* selects accusative, in the Icelandic example (8a) the verb *hvelfa* selects dative.

Zero subjects specified as {+ animate} occur in active or passive structures, cf. (7b), (8b) in sentences denoting controlled processes. Their role can be straightforwardly identified as 'non-referential human Agent'. Ukrainian shows an across-the-voice synonymy of active and passive constructions with a zero {+ animate} Agent, cf. (9a–b).

(9) Ukr. a. $\varnothing^{3Pl}$ {+ animate} Oficeriv$_{Acc.Pl.}$ zal'aka-l-y$_{3Pl}$.
'The officers were bullied,'
b. $\varnothing^{3Sg}$ {+ animate} Oficeriv$_{Acc.Pl.}$ bul-o3$^{Sg.N.Pst}$ zal'aka-n-o$_{Prt.Pst.3Sg.N.}$
'The officers were bullied', lit. '<it> was bullied to the officers.'

Semantic restrictions on the class of verbs licensing transitive impersonals with a {− animate} zero subject argument are language-specific. Russian does not allow

transitive impersonals by those causative predicates which require a {+animate} Causer, cf. *zapugat'* 'to intimidate', 'to bully', **ego*$_{3Acc.Sg}$ *zapugalo*$_{3Sg.N.Pst}$[15]. Somewhere transitive impersonals are licensed not by the lexical semantics of the verb alone, but by the event structure of the sentence. E.g., with *napugat'* 'to frighten smb', which is a semantic causative from *napugat'sa*[16], a sentence like *?mal'čika*$_{Acc.Sg}$ napuga-l-o$_{3Sg.N.Pst}$ ***vspyškami***$_{Instr.Pl}$ ***molnii*** 'the boy was frightened **by the lightning**' is much better than the **mal'čika*$_{Acc.Sg}$ *napuga-l-o*$_{3Sg.N.Pst}$ ***igruškoj***$_{Instr.Sg}$ 'the boy was frightened **by a toy**': the reason is that a sub-event 'impact of a lighting' more easily contributes to the resulting event 'situation P had a frightening effect over a boy' than a sub-event 'impact of a toy'. There are cases where lexical semantics of a verb is in conflict with general restrictions imposed by a zero subject construction. For instance, Russ. *zadolbat'* 'to cow smb.', 'to get at smb.' selects overt {+ animate} subjects while transitive impersonals in the 3Sg form denote processes not controlled by any human Agent. The judgements of native speakers whether they accept example (10) are split.

(10) Colloq. Rus. $^{?}\varnothing^{3Sg}$ {– animate} **Nas**$_{Acc.Pl}$ zadolba-l-o$_{3Sg.N.Pst}$. (a protest motto).
'We've got at', lit. '**to-us** was.cowed.'

Note that the anomaly in (10) again arises because the contribution of the dedicated sub-event 'Activity of some human Agents' to the resulting event 'Uncontrolled situation P that has an impact on the Patient' is semantically non-standard. At the same time, causatives from psych verbs which select a Patient {+ animate} argument are unproblematic regards their event structure since they just fix an impact of some factor X on Y's state of mind and do not specify whether the impact of X upon Y is caused by any intentional activity of human Agent. Let us examine *razdražat'* 'annoy', 'drive mad', 'get on one's nerves', which can be analyzed as semantic causative from an intransitive psych verb *razdražat'sa* 'to be annoyed', 'to be irritated'. Verbs from the *razdražat'* group do not license transitive impersonals in Russian, but they do license sententional subjects, cf. (2a) above which confirms that their subject argument is not specified as {+ animate}. Let us repeat the example (2a) below as (11).

(11) (11) Rus. Vas'u$_{Acc}$ razdražaet$_{3Sg}$ [$_{CP}$ čto Katia postojanno opazdyvaet].
'it gets on Vasja's nerves [$_{CP}$ that Kate always comes late].'

(11') Sub-event 'Kate's late arrivals' is part of the situation P 'factor X annoys Y.'

---

[15] Rus. *Zapugat'* 'to bully', unlike *napugat'* 'to frighten smb' seems to select only {+ animate} subjects, a restrictive condition that does not coincide with the ban on sententional subjects, since non-sententional {– animate} subjects are equally bad: ***sokraščenie zarplaty** zapugalo Vasju *'**Salary cuts** bullied Vasja' is ill-formed, while **sokraščenie zarplaty** napugalo Vasju '**Salary cuts** frightened Vasja' is OK.

[16] As for word formation, the vector is different, since the intransitive *napugat'sa* is derived from the causative *napugat'*.

Assume that K. is intentionally driving V. mad with her late arrivals. Still, from the viewpoint of Russian grammar and lexicon, V.'s irritation as an independent event not triggered directly by K.'s malicious attempts to irritate him. Standard causatives from non-psych verbs like *vynudit'* 'to force sm. to do smth', *zastavit'* 'make smb do smth', *sklonit' k* 'to dispose smb to smth' license an {+animate} zero subject controlling the 3Pl form which seems a more or less general feature of all Russian verbs, cf. (12a) and a silent argument with an approximate meaning 'situation P', cf. example (12b), with a lexicalized past participle *vynužden* 'forced'[17]. They also license overt sentential subjects as non-human Causers, as illustrated by (12c). In this case the CP filling in the subject position must be headed by a correlative pronoun $to_{3Sg.N}$ controlling the agreement form of the verb. Note that merging of bare that-clauses into subject position with causatives from the group *vynudit'* is impossible, cf. (12d).

(12) Rus. a. $\varnothing^{3Pl}$ {+ animate} **Ego**$_{3M.Acc.Sg}$ vynudi-l-i$_{3Pl.Pst}$ [$_{IP}$ uvolit'sa$_{Inf}$ s raboty].
'**He** was forced to quit his position <due to activities of some human Agents>'
b. On$_{3M.Nom.Sg}$. byl$_{3Sg.M.}$ vynužden [$_{IP}$ uvolit'sa$_{Inf}$ s raboty].
'he was forced to quit his position <due to some external circumstances or personal problems>'
c. [*To*$_{3Sg.N.,}$ [$_{CP}$ čto boss srezal emu zarplatu]] vynudi-l-o$_{3Sg.N.Pst}$ **ego**$_{3M.Acc.Sg}$ [$_{IP}$ uvolit'sa$_{Inf}$ s raboty].
'[$_{CP}$ that the boss cut down his salary] forced **him** to quit his position'.
d. *[$_{CP}$ čto boss srezal emu zarplatu]] vynudi-l-o$_{3Sg.N.Pst}$ **ego**$_{3M.Acc.Sg}$ [$_{IP}$ uvolit'sa$_{Inf}$ s raboty].

We have shown that Russian causatives verbs license sententional arguments and select overt subjects which are not specified as {+ animate}. This is explained by the event structure of causatives: {C} causes Y make P, where Y is specified as {+ animate} and factor C (Causer) may but not necessarily arises due to intentional activity of some {+ animate} X. Hence, Y may be forced to make P (say, quit one's position) both if some X aims at forcing him to do that and if factor C arises due to some other process (say, the communists came to power in Y-s country or Y suffers from severe depression). The presence of an {+ animate} Causer is only a sub-event of the causative situation C. At the same time, causative verbs are selective in taking zero subjects with the role of Agent—they license {+ animate;—referential} zero Agents, cf. (12a), but not {+ animate;—referential} zero Agents –*$\varnothing^{3Sg}$ {– animate} *mal'čika sil'no napugalo*, *$\varnothing^{3Sg}$ {– animate} *Vasju vynudilo opazdat'*, *$\varnothing^{3Sg}$ {– animate} *Vasju razdražalo opazdanijami* etc. These restrictions are likely explained by the fact that transitive impersonals normally describe uncontrolled situations as a whole and do not specify sub-events linked to their active participants. Deviations from this principle, as we have shown, lead to non-standard event structures and generate sentences not generally accepted by all speakers, cf. Rus. ?$\varnothing^{3Sg}$

[17] Rus. *vynužden* is morphologically a past participle and projects an event structure with an Agent argument, but the construction *X byl vynužden* does not classify with actional passives and an overt deagentive NP is strictly impossible: **X byl vynužden* ***Y-m*** *sdelat' Z*, intended 'X has been forced by Y to do Z'.

{– animate} *nas zadolbalo*, ??∅^{3Sg} {– animate} *mal'čika napugalo vspyškami molnii*. As for the ∅^{3Pl} {+ animate} zero subjects controlling the plural agreement on the predicate, they are licensed by causative verbs, since their event structure does not exclude, though does not require sub-events linked to a human Agent. A sentence like ∅^{3Pl} {+ animate} ***Ego***$_{3M.Acc.Sg}$ *vynudi-l-i*$_{3Pl.Pst}$ [$_{IP}$ *uvolit'sa s raboty*] asserts that some non-referential human Agents are responsible for Y's decision to quit his position, but does not imply that the activity of these human Agents was a sufficient condition for Y's act. Finally, we have shown a relevant distinction within the causative class which falls into two groups—causatives from non-psych verbs (*vynudit', sklonit' k*) which do not specify the structure of the causative situation P, and causatives from psych verbs (*razdražat', napugat'*) which specify that the causative situation is an affect or a mental reaction of an {+ animate} Causee Y. Causatives from non-psych verbs only take headed that-clauses (*to, čto P*-clauses) as sententional subjects and ban bare that-clauses (*čto P*-clauses), while causative from psych verbs license both headed and bare that-clauses as surface subjects.

In the last section of this paper I prove that the test on *to, čto P*-clauses is diagnostic in Russian for a wider class of propositional predicates and that the ability / inability of taking *to, čto P*-clauses and vs bare *čto P*-clauses hangs on the factive vs non-factive opposition and event structure.

	Causatives from non-psych verbs: *vynudit', sklonit' k*	**Causatives from psych verbs: *razdražat', napugat'***
headed *to, čto P*-clause as subject	+	+
bare *čto P*-clause as subject	–	+
{+ animate} Causee	+	+
{+ animate} Causer	±	±
dedicated sub-event linked with an active participant	+	–/?
transitive impersonals	–	–/?
zero {+ animate} subject	+	+

Fig. 1. Two groups of Russian causative verbs

5. Semantic Classes of Propositional Verbs and Syntactic Diagnostics

The notion of factive verbs was introduced in [Kiparsky 1970] and developed by Vendler (1980), Karttunen (1977), Arutyunova (1988), Padučeva (1986; 2004: 259), Bulygina & Šmelev (1988), Anna Zalizniak (2006) and others. The original idea was that verbs with a propositional argument split into two non-intersecting classes: verbs of knowledge and emotion from the first group (*know/ regret/ be glad that P)* bring about a presupposition that P is true and has a value of fact, while verbs of belief and speech from another class (*believe, tell, say that P*) do not bring about a factive

presupposition. Later research proved that non-factive verbs are heterogeneous and many verbs are used both in factive and non-factive contexts (cf. Eng. *tell*), so one must look for diagnostic contexts for all semantic classes, even though such tests as ability to lead indirect wh-questions *(X knows how Y did it* vs **X believes how Y did it*) do not cover all factive predicates, cf. [Bulygina & Šmelev 1988: 57–60] and one may need many tests for every language. An exact definition of non-factive verbs is a matter of discussion. Following Arutynova (1988), Padučeva (1986) and Zalizniak (2006: 449), I assume that the notion of 'situation' fits best to the propositional argument of non-factive verbs. Situations, i.e. arguments of such propositional attitudes as opinion, belief, evaluation etc. are intentional objects, inner states of mind, pictures, Gestalts, they are opposed to facts, i.e. propositions with the status of logical truth.

I argue that the syntax of *to, čto P*-clauses and *bare čto P* gives a clue for the description of Russian propositional predicates. The predictions are that a) if a predicate only licenses *to, čto P*-clauses, but not bare *čto P*-clauses as syntactic subjects it is factive, b) if a predicate only licensed *čto P*-clauses but not headed *to, čto P*-clauses, it is non-factive, c) if a predicate licenses both headed and bare *čto P* –clauses, it is ambivalent and its argument can be arranged both as fact and as situation. The *to, čto* test has been discussed earlier, but not in the version proposed in this paper where it is combined with analysis of sentence structure. Arutyunova (1988: 153) discusses *to, čto*—paraphrases like *Ivan uexal* 'Ivan left' → *to, čto Ivan uexal, rasstroilo menja* 'That Ivan left disturbed me' in the same context as full nominalizations like *Ivan uexal* → ***tot fakt**, čto Ivan uexal, rasstroil menja* '**The fact** that Ivan left disturbed me'. Padučeva (1986: 27) lists non-factive contexts where *to, čto*-clauses introduce a proposition with the status of 'situation', not fact, but concentrates on oblique forms of the correlative pronoun *to* where it is lexically governed by a preposition or a verb: *proizojti **iz-za togo**, čto P 'to happen **because of** P', načinat'sa **s togo**, čto P 'to begin **with** P', svodit'sa **k tomu**, čto P 'to amount to P'.* On the contrary, I focus on the uses of *to, čto*–clauses in the surface subject (structural Nominative case) or direct object (structural Accusative case) positions where *to* is <u>not</u> lexically governed by the matrix verb.

The form *to* is morphologically ambiguous between Nom.Sg. and Acc.Sg. The *to, čto*-clauses are syntactic nominalizations. Filling in the surface subject position, they impose a default agreement pattern in 3Sg. (in the past tense—3Sg.N. with nominal predicates and a past tense auxiliary), just as bare *čto*-clauses do, cf. (11) and (12c). It is however not clear beforehand whether *to, čto*-clauses agree with nominal predicatives. For the first, there is no evidence that CP-arguments of DPS predicatives (*mne stydno/ protivno/dosadno, čto P*) are raised to the surface subject position if overt dative subjects are present: this is possible if hierarchy (ii) holds for Russian, but impossible if hierarchy (i) is true. For the second, some DPS predicatives, cf. *stydno* 'it is a shame'*, žal'* 'it is a pity', *vse ravno* 'it is all the same'*, tak i nado* 'way to go' lack any agreeing counterparts in Modern Russian. Exactly these forms and a large group of other DPS predicatives, cf. *protivno* 'it is disgusting'*, dosadno* 'it is vexing'*, obidno* 'it is annoying' which retain counterparts in agreeing adjectives (*protivnyj, dosadnyj, obidnyj*) do not license *to, čto*-clauses. Meanwhile, predicatives from another group, cf. *izvestno* 'it is known', *stranno* 'it is strange', *bezrazlično* 'it does not matter' license both *to, čto*-clauses and bare *čto*-clauses. The analysis has shown that even if a predicative does not licence a *to, čto*-clause in DPS, the

corresponding agreeing adjective still may license a Dative-Nominative-Structure will full-fledged number-and-gender agreement with an NP *tot fakt* 'that fact', cf. (14b) and (14c).

(13) Rus. a. Mne$_{1Dat.Sg}$ bylo$_{3Sg.N.Pst}$ **stydno**$_{Pred}$, [$_{CP}$ čto tak vyšlo].
'I was ashamed [$_{CP}$ that it happened so].', lit. 'to-me was **shameful** that…'
b. *Mne$_{1Dat.Sg}$ bylo$_{3Sg.N.Pst}$ **stydno**$_{Pred}$, [*to* [$_{CP}$ čto tak vyšlo]].
c. *Mne$_{1Dat.Sg}$ byl$_{3Sg.N.Pst}$ **stydno**$_{Pred}$, /*__styden__$_{Adj,Nom.Sg.M}$ *tot*$_{Dem.Nom.Sg.M}$*fakt*$_{Nom.Sg.M}$ [$_{CP}$ čto tak vyšlo]]
int. 'I found the fact that it happened so shameful'.

(14) Rus. a. Mne$_{1Dat.Sg}$ bylo$_{3Sg.N.Pst}$ **protivno**$_{Pred}$, [$_{CP}$ čto tak vyšlo].
'I was disgusted [$_{CP}$ that it happened so].', lit. 'to-me was **disgusting** that…'
b. ??Mne$_{1Dat.Sg}$ bylo$_{3Sg.N.Pst}$ **protivno**$_{Pred}$, [*to* [$_{CP}$ čto tak vyšlo]].
c. Mne$_{1Dat.Sg}$ byl$_{3Sg.M.Pst}$ **protiven**$_{Adj,Nom.Sg.M}$ [$_{NP}$ ***tot***$_{Dem.Nom.Sg.M}$***fakt***$_{Nom.Sg.M}$ [$_{CP}$ čto tak vyšlo]].
'I found the fact that it happened so disgusting'.

(15) Rus. a. Mne$_{1Dat.Sg}$ bylo$_{3Sg.N.Pst}$ **bezrazlično**$_{Pred}$, [$_{CP}$ čto vse tak vyšlo].
'it was all the same to me that it happened so', lit. 'to-me was indifferent that…'
b. Mne$_{1Dat.Sg}$ bylo$_{3Sg.N.Pst}$ **bezrazlično**$_{Pred}$, [*to* [$_{CP}$ čto tak vyšlo]].
'the same.'
c. Mne$_{1Dat.Sg}$ byl$_{3Sg.M.Pst}$ **bezrazličen**$_{Adj.Nom.Sg.M}$ [$_{NP}$ ***tot***$_{Dem.Nom.Sg.M}$***fakt***$_{Nom.Sg.M}$ [$_{CP}$ čto tak vyšlo]].
'I was indifferent to the fact that it happened so'.

I conclude that a) that-clauses headed by *tot fakt*, *čto P* and *to, čto P* have different syntax, the correlative pronoun *to* does not stand in the same position as NP *tot fakt* and can only have default agreement with the predicate, b) DPS predicatives should not be mingled with agreeing adjectives, CP arguments of DPS predicatives only have default agreement.

Let us see the results of the *to, čto* P / bare *čto* P test in Russian. The predicates are classified in four groups—factive verbs of knowledge (class 1), causatives (classes 2 and 3), DPS predicatives with a CP-argument (classes 4 and 5) and non-factive verbs of believe/inner vision (classes 6 and 7). Propositional verbs taking nominative and non-nominative subjects are placed in different slots: CP-arguments (*to, čto* P / bare *čto* P) appear either in the direct object or in the subject position. For the sake of simplicity, I assume that with DPS predicatives the CP-argument is always in object position, if an overt DPS {+ animate} dative subject is present[18]. Russian factive verbs of knowledge always place CP-arguments in the object position (1), while non-factive verbs split into a 'personal' (6) and impersonal groups (7). Causative verbs only take CP-arguments as subjects (6, 7), while DPS predicatives, given the assumption above take them as objects (4,5).

[18] This stipulation is not essential for the analysis, since both subject CPs and direct object CPs stand in positions which are not lexically governed by the matrix predicate / preposition.

	Factive verbs	Causatives		DPS predicatives with a CP argument		Non-factive verbs	
	1	2	3	4	5	6	7
	+ Nom. subject: *X znaet, čto P/ to čto P*	From non-psych verbs: *To, čto P, vynuždaet X-a delatj Z*	From psych-verbs: *čto P/To,čto, P razdražaet X-a.*	'Factive group': *X-u izvestno/ važno, bezrazlično to, čto P/čto P*	'Non-factive group': *X-u stydno/ protivno, dosadno, čto P*	+ Nom. subject: *X dumaet, sčitaet, čto P*	− Nom. subject: *X-u mereščitsa, čto P*
to, čto P-clauses as subject/ direct object	+	+	+	+	−	−	−
bare *čto P*-clauses as subject/ direct object	+	−	+	+	+	+	+

Fig. 2. Predicate classes and the *to, čto*-clauses in Russian

The results can be interpreted in the following way. Factivity and capacity of taking *to, čto*-clauses as subject / direct object are related but independent values. *To, čto*-clauses are licensed by predicates projecting an event structure with a dedicated sub-event. This feature naturally combines with factivity. If a proposition has the status of fact, parts of it can easily be singled-out and highlighted: if *p* and *q* are sub-events of a fact P, then contrastive utterances that X knows *p* <but not necessarily knows *q*> and the corresponding prosodic cues for marking logical contrast [Yanko 1997: 209] are appropriate. If, on the contrary, proposition *p* has the status of an intentional object, situation, parts of it usually cannot be singled out, and there is no dedicated sub-event. Therefore non-factive predicates normally do not license *to, čto*-clauses. The ban on bare *čto P*-clauses with causatives from non-psych verbs (class 2) indicates that though causatives of this type license sententional subjects, the propositional argument has the status of fact and cannot be 'intensionalized'. This condition does not hold for causatives from psych-verbs (class 3): they subcategorize for {+ animate} Causees and neither ban nor require bare *čto P*-clauses. DPS predicatives split into a strictly non-factive class (5) that requires bare *čto P*-clauses and rules out *to, čto*-clauses, just as non-factive verbs (classes 6, 7) do, and ambivalent class (4), the members of which—cf. *izvestno/važno, bezrazlično* verbs behave exactly as causatives from non-psych verbs and license both bare *čto P*-clauses and *to, čto*-clauses: *mne važno, čto P ~ mne važno, čto P* 'it is important to me that P'.

Our interpretation could be undermined by non-factive verbs licensing *to, čto*-clauses. Such verbs exist in Russian, cf. the personal construction $ja_{\text{1Nom.Sg}}$ $verju_{\text{1Sg}}$ $\boldsymbol{v}_{\text{Prep}}$ $\boldsymbol{to}_{\text{Acc}}$, *čto* P, lit. 'I believe in that [$_{\text{CP}}$ that P]' ~ ja_{Nom} $verju_{\text{1Sg}}$, *čto* P and the impersonal construction $mne_{\text{1Dat.Sg.}}$ $veritsa_{\text{3Sg}}$ v_{Prep} $\boldsymbol{to}_{\text{Acc}}$, *čto* P, lit. 'me believes in that [$_{\text{CP}}$ that P]' ~ ja_{Nom} $verju_{\text{1Sg}}$, *čto* P. Yet neither the personal verb *veritj* nor the impersonal verb *veritsja* allow *to, čto*- clauses as subjects / direct objects, cf. (16a–b), so the test remains valid.

(16) Rus. a. Mne$_{1DatSg}$ ne$_{Neg}$ veritsa$_{3Sg}$ *to$_{Nom.Sg}$, [$_{CP}$ čto dannaja problema rešena][19].
'I hardly believe [$_{CP}$ that this problem is solved].'
b. Ja$_{1Nom.Sg}$ ne$_{Neg}$ verju$_{1Sg}$ *to$_{Acc.Sg}$, [$_{CP}$ čto dannaja problema rešena].
'I do not believe [$_{CP}$ that this problem is solved].'

A final point to be made is that expletive *eto*, unlike expletive *to*, is not selective to the semantic type of proposition and combines with some DPS predicatives from the *stydno, dosadno class* (5) which do not licence headed *to*-clauses.

(17) Rus. a. *Eto*$_{Expl}$ i dosadno$_{Pred}$, [$_{CP}$ čto dannaja problema ne rešena].
'It is but vexing [$_{CP}$ that this problem is not solved].'
b. [$_{CP}$ čto dannaja problema ne rešena], *eto*$_{Expl}$ i dosadno$_{Pred}$,
'the same,'
c.*dosadno *to*$_{Expl}$ [$_{CP}$ čto dannaja problema ne rešena].

6. Conclusions

Sententional complements in Russian and other languages with a nominative-accusative sentence patterns in most cases are internal arguments that can be raised to surface subject position where they alternate with oblique or expletive subjects, if a language has these kinds of sentence categories. Meanwhile, Russian causatives from non-psych verbs project an event structure where the sententional subject can be analyzed as Causer or even as Agent. The uses of correlative *to, čto P*-clauses in the positions of surface subject and direct object serve as diagnostics for factive predicates in Russian. Licensing of *to, čto P*-clauses hangs on a feature closely related to factivity—capacity of projecting an event structure with a dedicated sub-event. Inability of licensing *to, čto P*-clauses proves that a propositional predicate is non-factive. Russian has expletive elements *eto* and *to* which have different syntax. Expletive *eto* belongs to the matrix clause and does not form a constituent with the CP it antecedes. It alternates with oblique dative subjects in the surface subject position, can be separated from the correlative complement clause, has a propensity for fronting in its clause and is not sensitive to the semantics of DPS predicatives. Expletive *to* forms a constituent with its CP, cannot be separated from it and does not combine with non-factive DPS predicatives. These features of Russian expletive elements resemble the syntax of Germanic expletives like Eng. *it*, Da, Sw., Norw. *det*, Ger. *es*, but there are no one-to-one correspondences between the languages.

[19] The insertion of overt dative subjects in structures with eto and CP is impossible as shown in section 3.

References

1. *Apresian J. D.* 1985. Sintaksičeskie priznaki leksem // Russian Linguistics, Vol. 9, No. 2–3, 1985.
2. *Arutyunova N. D.* 1988. Tipy jazykovyx značenij. Ocenka. Sobytie. Fakt. Moscow: Nauka 1988.
3. *Belletti A. & Rizzi L.* 1988. Psych verbs and θ-theory. // Natural Language and Linguistic Theory, 6, 291–352.
4. *Bhatt R. M.* 1999. Verb movement and the syntax of Kashmiri. Dordrecht: Kluwer.
5. *Bulygina, T. V., Šmelev A. D.* 1988. Vopros o kosvennyx voprosax. Javl'aetsa li ustanovlennym faktom ix sviaz' s faktivnost'u? // Logičeskij analiz yazyka. Znanie i mnenie / N. D. Arutyunova (ed.). Moscow: Nauka, 46–63.
6. *De Cuba C., B. Ürögdi.* 2009. Eliminating factivity from syntax: sententional complements in Hungarian. // Approaches to Hungarian / M. den Dikken, R. M. Vago (eds.). Amsterdam: Benjamins, 29–64.
7. *Ekerot L.-E.* 1995. Ordföjld. Tempus, Bestämthet. Lund: Gleerups.
8. *Karttunen, L.* 1977. Syntax and semantics of questions // Linguistics and philosophy, 1997, N. 1
9. *Kiparsky, P. and C. Fact* // Progress in linguistics / M. Bierwisch, K. Heidolph (ed.). The Hague, 1970.
10. *Kniazev M.* 2012. A theta-theoretic account of the distribution of sentential complements
11. The case of Russian *čto*-clauses. // Proceedings of ConSOLE XX, 2012, 105–129.
12. *Lavine J.* 2014. Anti-Burzio Predicates: From Russian and Ukrainian to Icelandic // Vestnik MGGU. Serija Philologia. 2014, No. 1.
13. *Letuchiy A.* 2014. Sintaksičeskie svoistva sentencial'nyx aktantov pri predikativax // Vestnik MGGU. Serija Philologia. 2014, No. 2.
14. *Nygaard, M.* 1906. Norrøn Syntax. Kristiania: Aschenhoug.
15. *Partee, B. H.* 1979. Subject and object in Modern English. New York and London: Garland Publishing.
16. *Perlmutter D., Postal P.* 1983. The Relation Succession Law // Studies in Relational Grammar / Ed. by D. M. Perlmutter. Chicago-London, 30–80.
17. *Peškovskij A. M.* 1938. Russkij sintaksis v nauchnom osvechenii. Moscow
18. *Sperber W.* 1972. Ist die Zustandskategorie eine fhr die Beschreibung der Grammatik slawischer Sprachen notwendinge Wortart // Zeitschrift fhr Slawistik, 1972, Bd. 17, S. 401–409.
19. *Švedova* et alii 1982. Grammatika russkogo literaturnogo yazyka, II/ Švedova N. Y. et al. (eds.) Moscow: Russian Academy of Sciences.
20. *Testelets Y.* 2001. Vvedenie v obščij sintaksis. Moscow.
21. *Vendler Z.* 1980. Telling the facts // Speech act theory and pragmatics / Ed. By J. Searle et al. Dordrecht, 1980.
22. *Zalizniak Anna A.* 2006. Mnogoznačnost' v jazyke i sposoby eje predstavlenija. Moscow: jazyki slavianskoj kultury.
23. *Padučeva E. V.* 1986. O referencii yazykovyx vyraženij s nepredmetnym značeniem // Naučno-texničeskaja informacija. Serija 2. 1986, No. 1

24. *Padučeva E. V. 2004. Dinamičeskie modeli v semantike leksiki. Moscow: jazyki slavianskoj kultury.*
25. *Yanko T.* 2000. Bytovanie i obladanie: konstrukcii s glagolom *byt'*. // N. D. Arutyunova, I. B. Levontina (eds.). Logiceskij analiz yazyka Yazyki prostranstv. Moscow: Jazyki slavyanskoj kultury, 198–211.
26. *Zimmerling, A.* 2002. Tipologičeskij sintaksis skandinavskix jazykov. Moscow: Jazyki slavyanskoj kultury.
27. *Zimmerling A.* 2009. Dative Subjects and Semi-Expletive Pronouns in Russian // Studies in Formal Slavic Phonology, Morphology, Syntax, Semantics and Discourse Structure /Gerhild Zybatow, Uwe Junghanns, Denisa Lenertova, Petr Biskup (eds.). Peter Lang, 2009, 253–268.
28. *Zimmerling A. V.* 2012. Nekanoniceskie podležaščie v russkom yazyke. // Ot znacenia k forme, ot formy k znaceniju. Sbornik statej v cestj 80-letia Alexandra Vladimirovica Bondarko. Moscow: Jazyki slavyanskoj kultury, 568–590.
29. *Zimmerling A.* 2013. Transitive impersonals in Slavic and Germanic: Zero subjects and Thematic Relations. // Dialogue 2013. Computational linguistics and intellectual technologies, vol. 12 (19), 723–737.
30. *Zimmerling A.* 2013a. Sistemy poriadka slov slavyanskix jazykov s tipologičeskom aspekte. Moscow: jazyki slavianskoj kultury.
31. *Zimmerling, A.* 2014. Parametr nulevogo podležaščego i členenie teksta. // In memoriam Alexander E. Kibrik. Sankt-Petersburg: Aletheia, 201–218.

От редакции

Трудно примириться с мыслью, что мы больше не увидим на «Диалоге» ободряюще-заинтересованной улыбки Ильи Сегаловича.

Всегда очень занятой, он появлялся на встречах Оргкомитета и Редсовета еще не вполне освободившимся от предыдущего дела, но за какие-то минуты переключался полностью на обсуждающиеся заботы конференции, выглядел необыкновенно сконцентрированным, одновременно и продуктивным, и эмоциональным. Ему было интересно все, в чем был вызов и новизна: соревнования, неожиданные темы круглых столов, технологические новинки.

Илья оказал большое влияние на эволюцию «Диалога». В том, что конференция стала более деловой, современной, более ориентированной на реальные задачи, есть его большая заслуга. Всегда с уважением относясь к мнению старожилов конференции, сам будучи старожилом, он умел не со стороны, а с позиций своего в «Диалоге» человека убеждать отказываться от милых старых привычек и вносить в «Диалог» важное новое. В особенности значительным было влияние Ильи на становление процесса рецензирования: всегда защищая новые конференционные технологии, он на самом деле думал не столько о «модернизации», сколько об объективизации результатов рецензирования. Он и сам был одним из лучших рецензентов, всегда абсолютно неформальным и вовлеченным в своей реакции на рецензируемую работу. У него был свой путь в компьютерной лингвистике, тесно связанный с тем, чем он профессионально занимался. Он одним из первых увидел и оценил значимость того, чем стал в XXI веке Интернет. Первым на «Диалоге» он призвал лингвистов к изучению новых социолингвистических процессов в Интернете. Теперь трудно уже поверить, что большая часть профессиональной аудитории «Диалога» поначалу отнеслась к таким призывам с большим скепсисом, считая Интернет чем-то вроде громадной свалки языковых отходов.

Многие новые плодотворные идеи, такие как статистический анализ больших данных и машинное обучение, находили в его лице эффективного проводника на «Диалоге».

Границы возможного пролегали у него как-то иначе, чем у других. Все разумное казалось осуществимым. Пока был Илья, мир был лучше.

ИЛЬЯ СЕГАЛОВИЧ И РАЗВИТИЕ ИДЕЙ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ В ЯНДЕКСЕ

Зеленков Ю. Г. (yuryz@yandex-team.ru),
Зобнин А. И. (alzobnin@yandex-team.ru),
Маслов М. Ю. (maslov@yandex-team.ru),
Титов В. А. (uht@yandex-team.ru)

Яндекс, Москва, Россия

В статье рассматриваются наиболее важные и интересные лингвистические проекты, в которых участвовал и которыми руководил Илья Сегалович (1964–2013), один из создателей поисковой системы Яндекс. Среди этих проектов: разработка морфологического анализа и синтеза русских слов, позволяющего обрабатывать «новые» слова, не включенные в словарь; снятие морфологической омонимии для русского языка с помощью нормализующих подстановок; практическая транскрипция иностранной собственной и нарицательной лексики; автоматическая расстановка ударений и анализ поэтических текстов; создание эффективных методов распознавания нечетких дубликатов для текстовых документов; разработка информационно-справочной системы «Национальный корпус русского языка» и др. Описываются ключевые идеи и подходы, связанные с поиском решений сложных лингвистических задач и рассказывается о роли Ильи в изобретении этих подходов и их дальнейшем развитии. Приводятся примеры нетривиальных лингвистических алгоритмов, созданных Ильей вместе с коллегами.

Ключевые слова: Илья Сегалович, Яндекс, морфология, практическая транскрипция, анализ стихов, нечеткие дубликаты

ILYA SEGALOVICH AND DEVELOPMENT OF IDEAS OF COMPUTATIONAL LINGUISTICS TO YANDEX

Zelenkov Yu. G. (yuryz@yandex-team.ru),
Zobnin A. I. (alzobnin@yandex-team.ru),
Maslov M. Yu. (maslov@yandex-team.ru),
Titov V. A. (uht@yandex-team.ru)

Yandex, Moscow, Russia

In the article the most important and interesting linguistic projects led by Ilya Segalovich (1964–2013) — one of the founders of the Yandex search engine — are considered. He also took part in their development. The following projects are among them. Development of the morphological analysis and synthesis of Russian words with a possibility of processing «new» words not included in the dictionary; solving the problem of morphological ambiguity for the Russian language with the help of normalizing substitutions; practical transcription of foreign, individual and common words; automatic positioning of stresses and the analysis of poetic texts; creation of efficient methods of recognizing fuzzy duplicates for textual documents; development of the information and require system «The National Corpus of Russian», etc. Key ideas and approaches connected with the searching of solutions to complicated linguistic problems are described, and Ilya's role in the invention of these approaches and their further development is stated. Examples of non-trivial linguistic algorithms developed by Ilya in collaboration with his colleagues are given.

Key words: Ilya Segalovich, Yandex, morphology, practical transcription, analysis of poems, fuzzy duplicates

Введение

Илья Сегалович (1964–2013) внес большой вклад практически во все основные поисковые технологии Яндекса. И все-таки именно компьютерная лингвистика всегда была для него предметом особого интереса, как исследовательского, так и практического.

Основной чертой Ильи, как IT-профессионала, была уникальная способность находить и четко формулировать наиболее важные задачи, которые необходимо решать в данный момент, и в области информационного поиска в целом, и в компьютерной лингвистике в частности. Предлагаемые им алгоритмы всегда отличались оригинальным подходом и основывались на энциклопедическом знании той предметной области, к которой они относились. Уровень образованности Ильи поражал. Он всегда был в курсе самых последних публикаций практически по любому вопросу. Причем это касалось не только материалов текущих конференций, но и книжных новинок (в первую очередь зарубежных).

По инициативе и под руководством Ильи был выполнен перевод на русский язык одной из самых популярных в мире книг «Введение в информационный поиск» Кристофера Д. Маннинга и др. [7]. Специально для этой книги Ильей, совместно с известными специалистами, была проведена большая работа по созданию, упорядочиванию и редактированию современного толкового терминологического словаря для предметной области информационного поиска. В России работа такого объема была выполнена впервые.

Он придумал и был на протяжении ряда лет основным организатором конкурсов «Интернет-математика» и «Класс», которые существенно стимулировали проведение у нас в стране научно-исследовательских и прикладных разработок, а также создание учебных курсов в области компьютерного анализа данных. Под его редакцией вышли также несколько сборников этих конкурсов [5, 6].

Илья был одним из вдохновителей и первых участников проекта «Национальный корпус русского языка» [9, 12]. Благодаря именно его инициативе компания Яндекс в начале 2000-х вложилась в этот проект как финансово, так и людьми, осуществив поддержку разработки силами Виталия Титова, Андрея Кондратьева, Андрея Аброскина и других и предоставив поисковый модуль Яндекс.Сервер. Система снятия омонимии [3] и новая система морфологического разбора Яндекса, в которой усовершенствована обработка несловарных слов, обучены на корпусе со снятой омонимией из НКРЯ.

Илья Сегалович является автором алгоритма открытого (т. е. позволяющего с высокой точностью обрабатывать не входящие в словарь слова) морфологического анализа и синтеза для нескольких языков — ключевого лингвистического инструмента поисковой технологии Яндекса [10].

Помимо успешного решения этой базовой задачи компьютерной лингвистики Илья принимал активное участие в целом ряде других проектов, как чисто лингвистических, так и общепоискового характера, но опирающихся на лингвистику. О некоторых наиболее важных и интересных из таких проектов и пойдет речь дальше.

1. Морфологический анализ и синтез

В 1995–1996 гг. разработчики подразделения «Аркадия» фирмы Comptek International под руководством Ильи делали программную оболочку для электронных научных изданий «Грибоедов» и «Информ-Норматив». Главной задачей было создание полнотекстовой поисковой системы, и мы участвовали в формировании поискового конкорданса, т. е. в данном случае списка нормальных форм всех слов, встречающихся в текстах корпусов.

При построении конкорданса обнаружилось, что для «Информ-Норматива» «стандартный» морфологический словарь Зализняка покрывает список словоформ корпуса всего лишь наполовину. Это было довольно неожиданно для нас. Ведь корпус текстов относительно невелик, грамматически корректен, поэтому мы не ожидали в нем большого лексического разнообразия. Стала понятной важность задачи морфологической обработки слов, не содержащихся в словаре.

В тот период Илья предложил алгоритм морфологического анализа и синтеза для таких слов [10]. Новизна его идеи в сравнении с [2] состояла в «синтетической» части алгоритма, т. е. в возможности генерации гипотез моделей словоизменения для новых слов. Гипотезы строились в формате системы ЭТАП [8].

Проблема в том, что обычно таких гипотез получалось больше одной и нужен был способ выбора лучшей. Мы попробовали использовать информацию из корпуса текстов «Информ-Норматива», сравнивая парадигмы гипотез. Т.е. для гипотезы строили список словоформ, встречающихся в корпусе текстов, и применяли две несложные эвристики — «включение парадигмы» и «наличие нормальной формы» [10]. Поскольку корпус текстов «Информ-Норматива» был небольшим (да и эвристики немудреными), они срабатывали всего примерно в 20 % случаев. Так что этот результат мог быть только небольшим подспорьем специалистам, окончательно формировавшим конкорданс.

Программа MyStem, первая версия которой была написана Ильей Сегаловичем и Виталием Титовым, умеет строить такие гипотетические разборы для слов, не входящих в словарь.

Возможность получить еще более хороший результат для обработки как словарных, так и «несловарных» слов — с приписыванием значимых весов гипотезам разбора –возникла позже, с появлением и достаточным наполнением корпуса со снятой омонимией НКРЯ. Мы планируем в ближайшее время выпустить очередную версию программы — MyStem 3.0, поддерживающую ранжирующую морфологию и контекстное снятие омонимии.

Процедура ранжирования разборов на основе корпуса НКРЯ является дальнейшим развитием идей, предложенных Ильей в более ранних версиях морфологического алгоритма.

Для осуществления такого ранжирования применяется машинное обучение по корпусу. Просто упорядочить список разборов — плохой вариант, так как корпус является достаточно разреженным, и не каждое слово в нем представлено. Поэтому мы вычисляем по корпусу следующие величины: частоту каждой морфологической схемы (парадигмы), а внутри схемы — частоты каждой основы и каждого окончания слова. Каждому разбору конкретного слова соответствует своя схема, точно определяющая границы основы и окончания. Мы предполагаем, что события «встретилась данная основа *stem*» и «встретилось данное окончание *flex*» слова *word* при фиксированной схеме разбора *scheme* независимы. Поэтому вероятность разбора слова *word* по схеме *scheme* можно определить по формуле Байеса:

$$P(scheme \mid word) = \frac{P(word \mid scheme)P(scheme)}{P(word)} = \frac{P(stem \mid scheme)P(flex \mid scheme)P(scheme)}{P(word)}.$$

Этот подход представляет собой наивный байесовский классификатор. Его преимущество состоит, во-первых, в том, чтоданные о частотах распределены по основам и окончаниям (и поэтому их можно эффективно упаковать), а во-вторых, в том, что разреженность корпуса и низкие частоты отдельных форм слов уже не представляют проблемы: эти данные будут сглажены за счет других форм с такой же основой и других слов с таким же окончанием внутри данной схемы. Кроме того, полученные частоты дополнительно сглаживаются (например, простейшим методом Лапласа).

Точность этой модели на русском языке достигает 95,9% (по тексту леммы), в то время как точность простого выбора леммы с самой частотной схемой (baseline) равна 90%. Обращаем внимание на то, что здесь не используется контекст слова.

2. Проблема снятия омонимии при автоматической обработке текстов

Снятие омонимии полезно во многих приложениях компьютерной лингвистики, в частности, в поисковых системах оно может повысить точность

обработки запросов и сократить объем хранимой в архивах информации. К сожалению, эффективных подходов к решению этой задачи для русского языка на тот момент (2003 г.) не существовало (да и сейчас положение не намного лучше), и поэтому Илья предложил разработать новый оригинальный алгоритм решения этой задачи, опирающийся на информационные возможности Яндекса [3].

Для целей поиска необходимо было в первую очередь научиться снимать неполную (или морфологическую) омонимию, при которой разные слова совпадают не во всех, а только в нескольких грамматических формах, т. е. являются омоформами (имеют разные леммы). Примеры: три, стекло, стих, стали, белка, пара, вина, кос и др. Тем не менее полученные результаты позволяли (с небольшими изменениями) применять данный алгоритм и к разрешению полной (или лексической) омонимии.

Было принято решение использовать машинное обучение с учителем, поскольку основной целью являлась максимизация точности получаемых результатов. В дальнейшем появилась идея расширить алгоритм возможностью обучения без учителя, так как это существенно облегчало его настройку и обобщение подхода на другие языки.

В качестве данных для обучения алгоритма сначала использовался аннотированный массив текстов с морфологической разметкой и снятой омонимией из проекта «Национальный корпус русского языка» [9]. Затем для повышения полноты обучающей выборки был создан дополнительный веб-корпус. Процедура отбора текстов для этого корпуса учитывала поисковые возможности Яндекса и была автоматизирована с помощью специально разработанного генетического алгоритма, который обеспечивал оптимальную репрезентативность выборки документов из веба сразу по нескольким наиболее важным показателям (максимальное разнообразие явлений омонимии, жанров публикаций и их тематик).

При построении корпуса использовалась идея ранжирования самых частотных омонимов русского языка по степени «трудности выбора леммы». Омоним считался более «трудным», если для его разрешения с точностью выше фиксированного порога 0,96–0,97 требовался больший набор обучающих примеров для включения в корпус. Процедура ранжирования была полностью автоматизирована, и это позволило существенно (в разы) минимизировать размер обучающей выборки, не снижая качества получаемых результатов. В результате было выделено и ранжировано около 25 тыс. различных омонимов вместе с их контекстами.

Далее было нужно решить важную задачу удобного представления контекстов — слов, окружающих омоним слева и справа в предложении. Илья предложил оригинальную идею (полностью оправдавшую себя в дальнейшем) записывать отдельные элементы контекста в виде нормализующих подстановок, указывающих, сколько букв нужно отнять у словоформы с конца и на что их заменить, чтобы получить лемму. У флективных языков (к которым принадлежит и русский) нормализующие подстановки являются достаточно универсальным средством выражения грамматических свойств, поэтому мы и решили использовать именно их в качестве элементов контекста омоформы.

Затем возникла проблема количественной оценки «силы влияния» элемента контекста на выбор нужной леммы в зависимости от того, слева или справа от омонима расположен данный элемент и на каком расстоянии (сосед, через слово, ...). Кроме того, нужно было решить, сколько таких элементов брать, чтобы и словарь был достаточно компактным, и точность не пострадала. Была придумана функция, которая позволяла упорядочить элементы контекста по степени их влияния на выбор леммы. Оказалось, что наибольшим влиянием обладает сосед слева, затем сосед справа, затем следующие два элемента слева и следующий сосед справа. Сила влияния остальных слов резко падала, и поэтому они отбрасывались, т. е. при построении словаря учитывались только пять окружающих слов в указанном порядке, выраженном весовыми коэффициентами. На основе описанных процедур был построен словарь контекстов объемом около 150 тыс. обучающих примеров.

Алгоритм прошел своеобразную закалку во время одного из конкурсов «Интернет-математика», когда ему пришлось участвовать в поединке с алгоритмом, предложенным одним из участников. На основе проведенных тестов и экспертных оценок алгоритм Яндекса одержал победу.

Точность алгоритма составляет порядка 0,96–0,97 и при необходимости легко может быть увеличена с помощью автоматизированного дообучения. В настоящее время алгоритм входит в состав леммера — основного лингвистического инструмента Яндекса, работающего с 17 языками и у истоков которого также стоял Илья.

3. Практическая транскрипция собственной и нарицательной лексики

Современный интернет, основанный на идеологии Веб 2.0, характеризуется большим числом новостных потоков, множеством социальных сетей, развитием блогосферы, цифровой картографии, увеличением доли аудио- и видеоконтента и т. п. Все это приводит к существенному возрастанию удельного веса имен собственных как в веб-документах, так и в поисковых запросах. Как правило, эти имена написаны на самых различных языках и обозначают личные имена людей, географические названия, бренды, музыкальные группы и их произведения, фильмы и т. п.

Успешная обработка имен собственных для целей поиска осложняется тем, что в их написании царит произвол: имена, обозначающие одни и те же сущности, часто имеют несколько вариантов написания, как русскими буквами, так и латинскими.

Для наведения порядка в этом хаосе Илья в 2008 г. предложил разработать универсальный алгоритм практической транскрипции (записи русскими буквами максимально близкого звучания иностранного имени) собственной лексики. Задача была довольно амбициозная. Во-первых, правила транскрипции должны были быть вероятностными и составляться автоматически, т. к. ручные подходы в принципе не могли охватить всего разнообразия языковых

явлений. Статистический характер правил позволял создавать многовариантные гипотезы и тем самым мог как-то справиться с проблемой неоднозначности. Во-вторых, алгоритм должен был уметь определять исходный язык имени (или несколько языков, упорядоченных по вероятности) независимо от того, кириллицей или латиницей записано имя. В-третьих, нужно было научиться делать транскрипцию и в обратную сторону — с русского на язык оригинала. И наконец, необходимо было (хотя бы частично) научиться транскрибировать и нарицательную лексику, поскольку многословные имена собственные довольно часто ее включают (Empire State Building).

Для придания спортивного драйва проекту Илья даже пригласил внешнего разработчика со своими конкурентными идеями, но тот не выдержал предложенного темпа и постепенно сошел с дистанции.

За основу алгоритма было взято машинное обучение с помощью транскрипционных билингв (и в отдельных случаях, мультилингв) — параллельной записи имени на двух и более языках. Необходимое число билингв (порядка 100 тыс.) было получено автоматически из архивов Яндекса и открытых веб-источников, прежде всего из Википедии. Билингвы были написаны на 17 языках и включали все основные типы имен (имена людей, топонимы, бренды, ...)

Было перепробовано много вариантов алгоритма транскрипции, но наиболее удачной (простой в реализации, независимой от языка и эффективной по результатам) оказалась идея «метода сегментов», которую Илья и выбрал для последующей реализации и которая удовлетворяла всем вышеперечисленным требованиям к алгоритму. Сегмент — это группа рядом стоящих гласных или согласных букв, выделение которой для любого языка (за исключением *тайского*) тривиально. Правила транскрипции отдельного сегмента можно также легко получить из обучающей выборки следующим образом.

Сначала выполняем побуквенное выравнивание билингвы: «s t a — t e m e n t» = «с т е й т — м е н т», затем собираем сегменты и убираем пропуски (если они есть) в левой части: «st a t e m e nt» = «ст ей т — м е нт», и наконец убираем пропуски в правой части, объединяя сегменты в левой: «st a tem e nt» = «ст ей тм е нт». На практике все происходит гораздо проще, т. к. в подавляющем большинстве случаев (до 95 %) сегменты уже находятся во взаимно-однозначном соответствии и выравниваются без проблем.

Далее полученные по всей выборке соответствия сегментов группируются вместе с соседними сегментами слева и справа (учет контекста повышает точность), подсчитываются их частоты и окончательно формулируются вероятностные правила «перевода». В процессе транскрипции исходное слово разбивается на сегменты с контекстами, для каждого сегмента выбираются все варианты «перевода», вероятности суммируются, и результаты ранжируются по убыванию сумм вероятностей. Подход полностью симметричен как в направлении от латиницы к кириллице, так и обратно.

Алгоритм практической транскрипции используется в общепоисковых задачах и музыкальном поиске. В 2013 г. он был с успехом применен в сервисе «Яндекс.Карты» для перевода мировой карты примерно с 40 языков (включая тайский с его весьма специфической графикой для записи гласных) на русский.

«Метод сегментов» может быть успешно применен и для распознавания языков. Так, в 2010 г. было проведено соревнование по определению языка документа между существовавшим в Яндексе алгоритмом и новым, созданным на основе сегментов и дополненным механизмом языковых сигнатур, позволяющим легко разделять такие языки, как, например, русский и болгарский или шведский и датский. По тестам для 31 языка со счетом 28:3 победил новый алгоритм.

4. Автоматическая расстановка ударений и определение размера стиха

Интерес к автоматическому анализу поэтических текстов возник у Ильи примерно в 2009 г., когда в Яндексе был запущен проект «Стихолюб», а НКРЯ к тому времени содержал уже достаточно представительный размеченный поэтический подкорпус. Но это все были, так сказать, пассивные методы работы со стихами, требующими участия людей, а Илью в первую очередь привлекала возможность компьютерного подхода к проблеме.

Так возникла идея создать автономную (независимую от других лингвистических инструментов типа морфологии) и не очень сложную процедуру, которая с хорошей точностью могла бы распознавать ритмические фрагменты в текстах, определять размер стиха, рифмовку и пр.

Первая проблема на этом пути — автоматическая расстановка ударений в любых (в том числе отсутствующих в словаре) словах, т. к. от ее успешного решения зависели все остальные этапы. Была использована идея корреляции между буквенными концами слов и позицией ударения. С помощью машинного обучения был построен словарь буквенных концов объемом около 300 тыс. единиц. Вероятность правильной расстановки ударения с помощью такого словаря составляла почти 0,99 для нарицательной лексики и немного меньше для имен собственных.

Затем была написана довольно точная процедура фонетической транскрипции, и можно было приступать к автоматическому анализу стихов. Для классического силлабо-тонического стихосложения и некоторых неурегулированных размеров (дольник, тактовик, …) проблема, как ни странно, оказалась не очень сложной. Были записаны структурные формулы для этих размеров и для анализируемого стиха (после расстановки ударений) проверялась «близость» к этим формулам с помощью особой метрики. Самая близкая формула и определяла искомый размер и клаузулу (без ошибок анализировалось, например, стихотворение В. Брюсова «Ночь»). Результаты были довольно точные, программа «понимала», что такое спондей, пиррихий и другие премудрости типа синкопы икта, правда похуже.

Для улучшения результатов дополнительно использовался анализ строфы как целого. Илья придумал специальную функцию, которая могла ранжировать различные варианты определения структуры строфы по их внутренней «симметрии» и тем самым повышать качество разбора отдельных стихов, входящих в строфу. Например, в отдельных случаях нужно было смещать стандартное ударение («и г`орьки мне, горьк`и твои упреки»). С помощью специальной

таблицы также достаточно надежно определялась схема рифмовки (с попытками, где нужно, замены «е» на «ё»).

На этапе тестирования с помощью созданной процедуры было обнаружено несколько десятков тысяч ошибок в поэтической разметке НКРЯ (в основном расстановка ударений и определение размера).

Илье очень понравились полученные результаты, и он хотел внедрить анализатор стихов в разные сервисы типа «Яндекс.Словари», но тогда не сложилось. Совсем недавно на основе этой процедуры был создан «Автопоэт» — программа, которая сочиняет «стихи» произвольной формы (онегинская строфа, шекспировский сонет, ...) из запросов пользователей [1] (а сам Илья еще в 2011 г. предлагал сочинять стихи из твитов).

5. Обнаружение нечетких дубликатов в веб-документах

Распознавание нечетких дубликатов актуально для большого количества современных веб-приложений: это и улучшение качества индекса и архивов поисковиков за счет удаления избыточной информации, и объединение новостных сообщений в сюжеты на основе их сходства по содержанию, и фильтрация спама (как почтового, так и поискового), и установление нарушений авторских прав при незаконном копировании информации, и ряд других.

К 2007 г. уже было разработано множество интересных алгоритмов решения этой задачи [4], включая алгоритм, в создании которого участвовал и сам Илья [14]. Но все эти алгоритмы, без исключения, имели плохие показатели полноты, примерно 0,50–0,60, хотя точность была высокой — порядка 0,90–0,95. Поэтому Илья и предложил заняться повышением полноты без потери точности.

Самый надежный путь для этого — попарное сравнение (например, с помощью расстояния Левенштейна) всех документов из архива поисковика — был невозможен чисто физически. В Яндексе уже тогда было более 1 млрд документов, и число сравнений получалось астрономическое, измеряемое квинтиллионами. Нужно было что-то придумать. И тогда родилась идея провести декомпозицию этой гигантской матрицы, разложив ее на множество мелких подматриц, внутри которых выполнение попарных сравнений уже некритично.

Алгоритм декомпозиции оказался довольно простым. Нужно было для каждых пяти соседних слов (т. н. «шинглов») каждого документа указать длину документа в словах. Затем объединить одинаковые шинглы в последовательности по возрастанию длин документов. Было установлено, что, если соседние длины отличаются больше, чем на 15 %, эти документы не могут быть дубликатами. Таким образом, с помощью этого простого соотношения цепочки документов распадались на небольшие группы, внутри которых прямым попарным сравнением с помощью редакционного расстояния находились полудубликаты. Эксперимент показал 99%-ную точность и практически 100%-ную полноту при вполне приемлемой производительности.

Для дополнительного уменьшения размеров групп и повышения производительности, помимо числа слов в документах, использовалось также число

предложений, вместо «шинглов» брались 3 самых длинных предложения, а вместо расстояния Левенштейна сравнивались 5 самых длинных слов. При этом скорость существенно возрастала, а точность и полнота практически не менялись.

Результаты экспериментов: на коллекции РОМИП — полнота = 96%, точность=95%, F1-мера=0,95; на почтовой коллекции — полнота = 99%, точность=95%, F1-мера=0,97.

6. Оценка качества архивов поисковиков и определение объема Рунета

Несмотря на то что этот проект в меньшей степени связан с компьютерной лингвистикой, чем предыдущие, рассказать о нем необходимо, поскольку здесь ярко проявилась креативность Ильи, его способность к нестандартным подходам. В 2003–2004 гг. начался быстрый рост Рунета, и одновременно обострилась конкуренция между ведущими поисковиками. Для мониторинга этих процессов были необходимы автоматические процедуры, которые на регулярной основе могли бы измерять как темпы прироста русскоязычного сегмента Сети, так и основные количественные и качественные показатели поисковых машин.

Для оценки объема Рунета и определения величины русскоязычных архивов крупнейших поисковиков Илья оригинально модифицировал уже известный в то время подход Бхарата-Бродера [13], придав ему простоту и наглядность [11]. При определении доли выдачи какого-либо поисковика в выдаче других систем (основной пункт алгоритма Бхарата-Бродера) Илья предложил использовать небольшой массив из 120 редких однословных запросов, для которых выдача Яндекса не превышала 500 документов. Предполагалось, что в этом случае поисковики отключают все дополнительные фильтры и выдают полный набор документов из своего архива. Запросы отправлялись ко всем крупным поисковым системам, а результаты сравнивались с выдачей Яндекса и друг с другом. Подход Ильи опирался на гипотезу (основанную на многолетнем опыте разработки системы Яндекс) о пропорциональной зависимости между полными объемами архивов поисковиков и репрезентативными выборками из этих архивов.

Оказалось, что доля каждой поисковой системы в выдаче других поисковиков примерно одинакова, что доказывало независимость алгоритмов построения индексов этими системами. Тогда для определения объема Рунета нужно было разделить реальный размер базы Яндекса на его долю в выдаче, а для определения размера архива любого другого поисковика требовалось умножить размер архива Яндекса на долю этого поисковика. Просто и красиво, а главное, довольно точно, как показали независимые измерения.

Для измерения качества архивов Илья предложил уже полностью свои, оригинальные показатели «чистоты» (доля дубликатов в выдаче) и «свежести» (доля устаревших ссылок в выдаче). При определении «чистоты» использовалась (также придуманная Ильей для борьбы с почтовым спамом [11]) весьма остроумная методика «логарифмических шинглов». С ее помощью среди

полученных по редким запросам документов находились дубликаты, процент которых и характеризовал «чистоту» выдачи. При вычислении «свежести» делалась попытка найти основу слова редкого запроса в выдачах поисковиков. Процент неудачных попыток и определял «свежесть» архивов.

Можно, кажется, без конца вспоминать о лингвистических задачах, которыми интересовался Илья. Здесь мысли и об автоматической расстановке знаков препинания, и об определении морфемной структуры многокоренных слов, и о «лингвистических» расстояниях между частями речи и падежами и многие, многие другие. Но, как говорили древние, *sed satis verborum est*.

Подводя итог, можно сказать, что Илья был и душой, и мозговым центром практически всех лингвистических технологий Яндекса. Проекты, которые он начинал и которыми руководил, стали к настоящему времени мощными инструментами автоматической обработки ЕЯ-текстов, помогающими Яндексу находиться в числе мировых лидеров в области информационных технологий.

Литература

1. Автопоэт Яндекса. http://blog.yandex.ru/post/73398/
2. *Белоногов Г. Г., Зеленков Ю. Г.* Алгоритм морфологического анализа русских слов. Вопросы информационной теории и практики, N 53, М., ВИНИТИ, 1985. — 156 с.
3. *Зеленков Ю., Сегалович И., Титов В.* Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог'2005». — М.: Наука, 2005. 616 с.
4. *Зеленков Ю., Сегалович И.* Сравнительный анализ методов определения нечетких дубликатов для Web-документов. // Труды девятой всероссийской научной конференции RCDL'2007. Переславль-Залесский, 15–18 октября 2007.
5. Интернет-математика 2005. Автоматическая обработка веб-данных. М., 2005. — 504 с.
6. Интернет-математика 2007. Сб. работ участников конкурса. Екатеринбург, Изд-во Урал. ун-та, 2007. — 224 с.
7. *Кристофер Д. Маннинг* и др. Введение в информационный поиск. М., Вильямс, 2011. — 528 с.
8. Лингвистическое обеспечение системы ЭТАП-2. М.,1989.
9. Национальный корпус русского языка. http://www.ruscorpora.ru/
10. *Сегалович И., Маслов М.* Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов. // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог'99». Т. 2. С. 547–552. Казань, 1998.
11. *Сегалович И., Зеленков Ю., Нагорнов Д.* Методы сравнительного анализа современных поисковых систем и определения объема Рунета. // Труды восьмой всероссийской научной конференции RCDL'2006. Суздаль, 17–19 октября 2006.

12. *Сичинава Д. В.* «Национальный корпус русского языка: очерк предыстории» (2005). http://www.ruscorpora.ru/sbornik2005/03sitch.pdf
13. *Bharat K. and Broder A.* A Technique for Measuring the Relative Size and Overlap of Public Web Search Engines. Proc. of the 7th International World Wide Web Conference, April 1998.
14. *Ilyinsky S., Kuzmin M., Melkov A., Segalovich I.* An efficient method to detect duplicates of Web documents with the use of inverted index. WWW'2002 — Eleventh International World Wide Web Conference.

Abstracts

AUTOMATIC CREATION OF HUMAN-ORIENTED TRANSLATION DICTIONARIES

Antonova A. (antonova@yandex-team.ru), **Misyurev A.** (misyurev@yandex-team.ru), Yandex, Moscow, Russia

This paper addresses the issue of automatic acquisition of a human-oriented translation dictionary from a large parallel corpus. Automatically generated dictionary entries can enrich the output of a statistical machine translation system. We describe an automatic approach to the extraction of translation equivalents, and dictionary entry construction: grouping of synonymic translations, selection of illustrative context examples. The extraction of possible translations is based on statistical machine translation methods. The selection of lemmatized and linguistically motivated phrases is done with the help of morpho-syntactic analysis. In contrast to human-built dictionaries, an automatic dictionary usually contains a certain amount of noisy translations, as a consequence of systematic alignment mistakes and corpus imperfections. A noise reduction approach is proposed. We also provide the result of an evaluation experiment and the comparison of frequency distribution of words in the queries to the dictionary and the frequency distribution of words in plain text.

IDIOMATIZATION AND GRAMMATICALICATION IN NON-STANDARD CONSTRUCTIONS

Apresjan V. Yu. (vapresyan@hse.ru), National Research University Higher School of Economics, Moscow, Russia

The paper is a corpus research of the Russian construction *wh-word + negative particle X,P* (as in *Kak ni trudno, nado starat'sja* 'However difficult, one has to try'; *Čto on ni prosil, vse emu davali* 'Whatever he asked for, he was given') as a typical representative of a certain class of syntactic objects, namely, non-standard constructions, which reveals the following properties: 1) only one or several lexemes ("favorites") account for up to a half of all encountered realizations; 2) non-standard constructions are non-compositional; 3) realizations with certain "favorites" result in idiomatization and grammaticalization of particular expressions which become separated from the "mother" construction. The choice of "favorites" is triggered by the process of mutual semantic attraction: the interaction of the construction semantics and the semantics of filler lexemes. This choice is also influenced by the linguistic worldview typical of a particular language.

AUTOMATIC ENRICHMENT OF INFORMAL ONTOLOGY BY ANALYZING A DOMAIN-SPECIFIC TEXT COLLECTION

Astrakhantsev N. A. (astrakhantsev@ispras.ru), **Fedorenko D. G.** (fedorenko@ispras.ru), **Turdakov D. Y.** (turdakov@ispras.ru), Institute for System Programming of the Russian Academy of Sciences, Moscow, Russia

The core part of an entity linking system, in particular one oriented to wikification, is ontology, which is often informal and supports semantic relatedness as the only type of relation. Most of these systems suffer from the problem of ontology incompleteness. It is especially important for specific domains, since often the only source of extractable knowledge is plain text. This paper formulates the incompleteness problem as a task of ontology enrichment from domain-specific texts and presents a novel approach that combines state-of-the-art methods for terminology enrichment, our own ML-based method for homonymy detection, and methods adopted from the related field for relations extraction. Experimental evaluation shows that the bottleneck is terminology enrichment step: its average precision is about 35%, which is inapplicable for automatic usage, especially taking into account the strict requirements for ontology correctness; however, recall is high enough to help semi-automatic terminology enrichment. We also show that the best features for terminology enrichment differ from those for classic terminology recognition task.

ACTIVITY OF PARTICIPANTS IN A CONVERSATION: METHODS OF LINGUISTIC ANALYSIS

Baranov A. N. (baranov_anatoly@hotmail.com), Vinogradov Russian Language Institute of the Russian Academy of Sciences, Russia

The paper deals with the phenomenon of activity of dialogue participants. Analysis of participants' activity in a conversation is of great importance for theoretical linguistics as well as for applied linguistics. In forensic linguistics, analysis of activity can be used as an objective parameter for the qualification of real communicative goals of participants. The paper introduces three major methods of analysis of the phenomenon discussed: 1) the method of **communicative activity**, i.e. the amount of illocutionary independent speech acts of a participant in a dialogue or in its relevant part; 2) the method of **thematic activity**, the analysis of which enables the detection of exactly which participants independently introduces the main themes in a conversation; 3) the method of **quantitative activity**, based on calculating the amount of words associated with a specific theme in a conversation. We discuss the different types of correlation between the three methods.

MULTIMODAL AND CROSS-MODAL DISTRIBUTIONAL SEMANTICS: TOWARDS COMMON SEMANTIC SPACE FOR WORDS AND THINGS

Baroni M. (marco.baroni@unitn.it), Center for Mind/Brain Sciences, University of Trento, Italy

Distributional semantic models (DSMs) capture various aspects of word meaning with vectors summarizing their patterns of co-occurrence in large text corpora, under the assumption that the contexts in which words occur are good cues of what they mean. DSMs have been very successful empirically, and they have been used to model increasingly sophisticated linguistic and cognitive phenomena.
However, current DSMs account for linguistic meaning entirely in terms of linguistic signs (the "meaning" of a word is a summary of the linguistic contexts in which the word occurs). This leads to two big conceptual problems: lack of grounding and lack of reference. Concerning the former, cognitive scientists have accumulated plenty of evidence that, for human beings, meaning is strongly embodied in the sensory-motor system, so a semantic theory that completely dissociates meaning from perception and action is, a priori, a rather implausible model of how humans work — a fact that has also empirical consequences in the surprisingly bad performance of DSMs on simple tasks requiring perceptual information. Lack of reference is perhaps an even more serious problem. A theory that has no way to connect the semantic representation of a linguistic expression to states of the world is clearly missing something fundamental about language, as it has no way to explain how we can talk about things!
Interestingly, in the last decade, it has become common in computer vision to represent images through vectors recording the distribution of automatically extracted discrete visual features in them — a representation that is very similar to the one that DSMs assume for words. This suggests that we might be able to free DSMs from their textual cage by establishing a connection with the visual world by means of such vector-based image-representation techniques.
In my talk, after a brief general introduction to distributional semantics, I will discuss experiments we carried out in the last few years in which we tackle the grounding problem (DSMs with richer multimodal semantic representations that combine linguistic and visual features), and recent work in which we started dealing with the reference issue (how to map images and linguistic expressions across modalities to a common space, in order to link language to the world out there). The case studies I will present include simulating human semantic similarity judgments, predicting the color of objects, modeling brain data and learning names and verbally-expressed attributes of objects present in pictures from indirect evidence.

VARIATIONAL CORPUS STATISTICS USING AUTHOR PROFILES

Belikov V. (vibelikov@gmail.com), RSUH, Moscow, Russia,
Kopylov N. (Nikolay_Ko@abbyy.com), RSUH, ABBYY, Moscow, Russia,
Selegey V. (Vladimir_S@abbyy.com), RSUH, ABBYY, Moscow, Russia,
Sharoff S. (s.sharoff@leeds.ac.uk), RSUH, Moscow, Russia; University of Leeds, UK

This paper is based on research carried out in the framework of our project on the General Internet Corpus of Russian (Geekrya) . The need to use large-scale corpora automatically collected from the Web was first recognized in computational linguistics. Recently, the lack of data in "manually-built" corpora led to recognition of the importance of Web-derived corpora in traditional linguistic research.

The principal difference of Geekrya from the two other large web corpora of Russian (RuWac and RuTenTen) is that the latter were produced by indiscriminate crawling of the Russian Internet, resulting in no metatext markup available for their data.
GEEKRYA is different since its contents is split into "segments" which we define as a compact set of webpages sharing a general communicative purpose expressed in text-rich content. We extracted information about the authors from their profiles when this was specified.
The total size of indexed Geekrya amounts to 12 billion words, the segments with known a priori metatext parameters are listed below (size given in millions of words).

Segment	Gender	Age	Region
blogs.mail.ru	164	81	113
livejournal.com	0	1,800	5,600
vk.com	2,000	1,600	1,600
news	0	0	0
magazines.russ.ru	258	0	0
forums (adw.ru)	163	0	0
Total:	2,585	3,481	7,313

The magazines.russ.ru segment, for example, contains all the texts from this resource (mostly published fiction and literary criticism). Author's gender has been extracted for 84.3% its texts, the size of the male subcorpus is—194 MW, the female one is 64 MW.
Information about the author' profiles within the individual segments helps in variational analysis. The paper lists several studies on the gender profiles of discourse words, collocations and idioms, as well as on the regional distribution, for example, comparing word uses in Siberia against the rest of the Russian-speaking world.

USING DISTRIBUTED REPRESENTATIONS FOR ASPECT-BASED SENTIMENT ANALYSIS

Blinov P. D. (blinoff.pavel@gmail.com), **Kotelnikov E. V.** (kotelnikov.ev@gmail.com), Vyatka State Humanities University, Kirov, Russia

The article is focused on aspect-based sentiment analysis, which is a specific version of the general sentiment analysis task. Its goal is to detect the opinions expressed in the text on the level of significant aspects of the specified entity. An overview of the existing approaches and previous work is presented.
The main result of our work is a new method of aspect-based sentiment analysis based on the distributed representations of words. Such representations are obtained by using deep learning algorithms. The method includes the well-known algorithm of training distributed representations of words, two new techniques for constructing the aspect and sentiment lexicons, and an algorithm for calculating aspect scores.
Examples of aspect and sentiment terms are given. The vectors of resulting terms are visualized using the t-SNE method. The article presents the results of experiments on a test corpus for three aspects—"food", "interior" and "service", which yield aF1-measure increase of 11 to 16% as compared to the baseline.

ONE OF THE MOST FREQUENT ITEMS IN RUSSIAN SPONTANEOUS SPEECH: *БЛИН* FROM LINGUISTIC AND SOCIOLINGUISTIC POINTS OF VIEW

Bogdanova-Beglarian N. V. (nvbogdanova_2005@mail.ru), Saint Petersburg State University, St. Petersburg, Russia

The paper is dedicated to some peculiar functions of one of the most frequent items in Russian spontaneous speech, "блин", which formally being a word, is in fact more of a functional item). Using a Russian speech corpus ('One Speech Day' sub-corpus) we explored the historical change of the item; from an interjectionally used euphemism for an extremely rude slang word meaning 'whore'—through an acceptable colloquialism—to an almost meaningless clitic. So the evolution of this word begins at the point of being absolutely unacceptable in everyday speech, continues through being common and existing in any kind of neutral speaking, and ends as an ornamental word that probably lost the connection with its first meaning completely.
The final item does not have any meaning, lacks grammar categories, is not marked by intonation and has almost no emotional connotation. Normally such words are mostly used by men; but in this particular case gender does not play any role.

ANAPHORA ANALYSIS BASED ON ABBYY COMPRENO LINGUISTIC TECHNOLOGIES

Bogdanov A. V. (abogdanov@abbyy.com), **Dzhumaev S. S.** (sdzhumaev@abbyy.com), **Skorinkin D. A.** (dskorinkin@abbyy.com), **Starostin A. S.** (astarostin@abbyy.com), ABBYY, Moscow, Russia

This paper presents an anaphora analysis system that was an entry for the Dialog 2014 anaphora analysis competition. The system is based on ABBYY Compreno linguistic technologies. For some of the tasks of this competition we used basic features of the Compreno technology, while others required building new rules and mechanisms or making adjustments to the existing ones. Below we briefly describe the mechanisms (both basic and new) that were used in our system for this competition.

THE DISCOURSE WORDS AND REFERENCE IN THE PROCESS OF UNDERSTANDING

Borisova E. G. (efcomconf@list.ru), Moscow City Teachers' Training University, Moscow, Russia

The article addresses some aspects of the process of understanding, namely the reference of the components of utterances. The referential activity of the Hearer is regarded as a part of his actions in analyzing the sentences. These actions of the Hearer can be corrected by the Speaker with the help of discourse markers, modal particles etc. These entities are no markers of a referential status of nouns, still they can help reveal this status in some complicated cases, as follows: A non-actualized (though definite) name is used as a topic of an utterance. The Russian particle—*to* that marks such topics can help reveal the definite status of the name. Some other complicated cases of topic formation can be marked by particles *vot* and *von*.
The word that denotes something known (maybe mentioned in the previous context) can be revealed with the help of the particles *vot*, *imenno*, *kak raz*. It concerns not only nouns but also predicates.
The indefinite status of a noun can be demonstrated with the help of the particle *tam*, which is used to denote unimportance of some fact or a noun.

ONTOLOGY AND INTEGRATION OF FORMAL AND LEXICAL SEMANTICS

Borschev V. B. (borschev@linguist.umass.edu), **Partee B. H.** (partee@linguist.umass.edu), University of Massachusetts, Amherst, USA

Formal and lexical semantics can be integrated if they speak the same language. We claim that a substantial part of lexical semantics can be incorporated into formal semantics without adding to the latter any new mechanisms. This talk continues the authors' work on the ontology and the semantics of measure constructions in Russian. The work concerns expressions like *dva stakana moloka, polkorziny gribov, tri meshka muki* (*two glasses of milk, half a basket of mushrooms, three bags of flour*), etc., describing various kinds of *containers*, or corresponding measures based on them, and their contents—*portions of substances*. In our previous works, describing ontological information, including *sorts* of things and the words and expressions that designate sorts, we did not include those sorts in our formal semantic analyses. We do that in the present work, declaring *sorts* as *types* and thereby significantly expanding Montague's system of types. On the one hand this gives us the means for specifying various aspects of the ontology, and on the other hand it lets us more fully specify the semantics of the constructions under consideration. The substantive goals of this research are, in part, to be able to describe and explain co-occurrence constraints and ideally to be able to formally distinguish well-formed from ill-formed expressions in this domain.

A VIRTUAL RUSSIAN SENSE TAGGED CORPUS AND CATCHING ERRORS IN A RUSSIAN ↔ SEMANTIC PIVOT DICTIONARY

Dikonov V. G. (dikonov@iitp.ru), IITP RAS, Moscow, Russia,
Poritski V. V. (v.poritski@gmail.com), BSU, Minsk, Belarus

There are areas in computational linguistics, where a word-sense tagged corpus becomes a necessary prerequisite or gives a significant boost to research. Unfortunately, publicly available corpora of this kind are extremely rare and making them from scratch is a very long and costly process. No corpus of Russian with unambiguous word-sense tags has been published so far. This paper describes an experimental approach of creating a virtual equivalent of a Russian sense tagged corpus and putting it to some real use. The virtual corpus was created using two public resources: the English SemCor corpus and our free multilingual semantic pivot dictionary, called the "Universal Dictionary of Concepts". The dictionary provides information sufficient to find sense-specific translations for nearly all sense-tagged words in SemCor. However, the pivot dictionary itself is under development and we are looking for the ways to improve it. We used the existing Russian volume of the pivot dictionary to calculate lexical context vectors for individual senses of 13,832 Russian words, supposedly equivalent to the vectors that could be obtained from a real Russian translation of SemCor. Another set of vectors representing real usage of the same Russian words was extracted from a medium-size corpus of Russian without any semantic markup. The vector similarity score proved to be a useful factor in judging the correctness of links between Russian words and word senses similar to ones registered in the Princeton Wordnet. It helped to rank over 21,000 of such links out of 56,000 known and significantly reduce the amount of the manual work required to proofread the dictionary.

DISCOURSE WORDS IN GENERAL QUESTIONS: RUSSIAN-GERMAN NEAR-EQUIVALENTS

Dobrovol'skij D. O. (dm-dbrv@yandex.ru), **Levontina I. B.** (irina.levontina@mail.ru),
Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

The paper discusses Russian discourse words such as *razve, neuzheli, chto, chto li, kak,* etc. Cf. *Ty chto / chto li / kak, s nami idesh'*? ≈ 'What about you, are you coming along?', and their German near-equivalent etwa (cf. *Gehst du etwa mit?*). Our data show that translatability and semantic equivalence are different phenomena. Both Russian and German possess a rich inventory of question particles, which makes it possible to find a suitable translation for nearly every utterance, even a translation containing a particle. However, this does not imply that the corresponding particles are semantically equivalent. The analysis shows that such particles, being functionally equivalent, i.e. interchangeable in particular utterances, display rather remote semantic resemblance. The German particle *etwa* is conceptually based on the idea of approximateness. That is why it weakens the illocutionary force of the utterance, whereas the Russian particles *chto, chto li, kak* directly appeal to the interlocutor and, therefore, reinforce the speaker's attitude. However, both German *etwa* and Russian *chto, chto li, kak* stress the speaker's involvement in the situation. This property determines their functional similarity.

MODALS AND THE SUBJUNCTIVE

Dobrushina N. R. (nina.dobrushina@gmail.com), National Research University Higher School of Economics, Moscow, Russia

I consider constructions that involve the modal verb *moch'* or the modal adjective *dolzhen* and the subjunctive particle *by*. I argue that, with respect to the subjunctive, these modals behave differently from regular verbs. Their subjunctive is often functionally identical to the indicative; in contexts where other verbs obligatorily take the subjunctive form, these two predicates may use the indicative. The main factor that controls omissibility of the subjunctive particle is shown to be an epistemic interpretation. I consider some typical cases where the subjunctive and the indicative are synonymous for these predicates, and those where they are not. Thus, in the apodosis of conditional constructions the particle is often omitted, although, in general, Russian prefers a symmetrical use of the subjunctive in both protasis and apodosis. On the other hand, when in the protasis, the particle is not omitted. The subjunctive is often used with the modals for pragmatic purposes, such as politeness. The paper is based on the data from the Russian National Corpus.

QUERY EXPANSION IN INFORMATION RETRIEVAL: WHAT CAN WE LEARN FROM A DEEP DNALYSIS OF QUERIES?

Ermakova L. M. (liana.ermakova@irit.fr), Institut de Recherche en Informatique de Toulouse, CNRS UMR 5505, Université de Toulouse, France; State National Research University, Perm, Russia, **Mothe J.** (josiane.mothe@irit.fr), Institut de Recherche en Informatique de Toulouse, CNRS UMR 5505, Université de Toulouse, France, **Ovchinnikova I. G.** (ira.ovchi@gmail.com), Perm State National Research University, Perm, Russia

Information retrieval aims at retrieving relevant documents answering a user's need expressed through a query. Users' queries are generally less than 3 words which make a correct answer really difficult. Automatic query expansion (QE) improves the precision on average even if it can decrease the results for some queries. We propose a new automatic QE method that estimates the importance of expansion candidate terms by the strength of their relation to the query terms. The method combines local analysis and global analysis of texts. We evaluate the method using international benchmark collections and measures. We found comparable results on average compared to the Bo2 method. However, we show that a deep analysis of initial and expanded queries brings interesting insights that could help future research in the domain.

WORKING MEMORY AND RUSSIAN LANGUAGE: FROM COMPREHENSION TO PRODUCTION

Fedorova O. V. (olga.fedorova@msu.ru), **Potanina Ju. D.** (binechka-paveletskaja@mail.ru), Lomonosov Moscow State University, Moscow, Russia

Working memory and long-term memory differ in many ways. One difference is in the storage capacity of each. Traditionally, the capacity of the working memory has been measured by a memory span task in which the individual hears series of items and must repeat them. Most of the research has focus on individual differences in working memory capacities. Daneman and Carpenter (1980) developed the Reading span test, which they interpret as providing a measure of an individual's working memory capacity. The subject is given a series of sentences to read, and then must recall the last word from each of the preceding sentences. Span is calculated as the maximum number of sentences on which the subject can perform this task perfectly. In 1986 Daneman and Green developed the Speaking span test. Most of the research has done on English-speaking individuals. The main goal of this paper is to provide and describe the Verbal span tests on Russian material. The present study shows how the use of the notion of verbal working memory contributes to our understanding the individual differences in language comprehension and language production mechanisms. Using Russian adaptations of the working memory reading span and speaking span tests we demonstrated that the working memory capacity is really correlated with some referential processes, as well as it is a predictor of verbal fluency.

RING AND *GRAPPOLO*: FINGERTIP CONNECTIONS IN RUSSIAN GESTICULATION AND THEIR MEANINGS

Grishina E. A. (rudi2007@yandex.ru), Institute of Russian Language RAS, Moscow, Russia

The study analyzes the main types of Russian gestures, which are based on the connection of one's fingertips (configuration *exactly, feather, bunch*). We distinguish five semantic groups, which correspond to these configurations ('exactness', 'small object', 'object', 'center', 'connection'). We also compare the linguistic functions of the fingertip connections and the hand physical contact.

TOWARDS A WORD SENSE FREQUENCY DICTIONARY

Iomdin B. L. (iomdin@ruslang.ru), Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia; National Research University Higher School of Economics, Moscow, Russia, **Lopukhina A. A.** (nastya-merk@yandex.ru), Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia, **Nosyrev G. V.** (grigorij-nosyrev@yandex.ru), Yandex, Moscow, Russia

Analyzing several Russian nouns denoting everyday life objects, we explain why a word sense frequency dictionary is necessary. Techniques of calculating the approximate frequencies are proposed, based on the analysis of native speaker surveys and the annotation of the most frequent collocations in a large text corpus (we used the huge RuTenTen11 corpus integrated into the Sketch Engine system). A word sense dictionary could be used in a variety of NLP tasks, in particular for a probabilistic word sense disambiguation without available context, in creating second language learning resources, as well as in academic lexicography. Besides, studies of sense sets of polysemous words and their comparative frequencies are important for the linguistic theory, because they shed light on the evolution of the lexical system.

VALENCIES OF RUSSIAN PREDICATE NOUNS AND MICROSYNTACTIC CONSTRUCTIONS

Iomdin L. L. (iomdin@iitp.ru), Kharkevich Institute for Information Transmission Problems of the Russian Academy of Sciences, Moscow, Russia, **Iomdin B. L.** (iomdin@ruslang.ru), Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia; National Research University Higher School of Economics, Moscow, Russia

The paper discusses valency realizations of Russian predicate nouns in certain types of syntactic constructions (mainly, existential ones like *Mne net neobxodimosti sdavat ekzamen* 'There is no need for me to take the exam'; lit. 'to me there is no necessity...') where these realizations are not directly linked with the nouns concerned. In these cases, subcategorization frames of nouns are insufficient to account for the correct semantic interpretation of the construction in text analysis, or the adequate choice of valency implementation in text generation. For every word, detailed information on how its valencies are implemented within particular constructions should be supplied in the dictionary.

THE IMPACT OF MORPHOLOGY PROCESSING QUALITY ON AUTOMATED ANAPHORA RESOLUTION FOR RUSSIAN

Ionov M. (m.ionov@corp.mail.ru), Mail.ru Group, Moscow, Russia, **Kutuzov A.** (andrey.kutuzov@corp.mail.ru), Mail.ru Group, National Research University Higher School of Economics, Moscow, Russia

The paper deals with the problems of creating and tuning a system of automated anaphora resolution for Russian. Such a system is introduced, combining rule-based and machine learning approaches. It shows F-measure from 0.51 to 0.59. Freeling serves as an underlying morphological layer and an account of its quality is given, with its influence on anaphora resolution workflow. The anaphora resolution system itself is available to download and use, coming with online demo.

DATA-DRIVEN METHODS FOR ANAPHORA RESOLUTION OF RUSSIAN TEXTS

Kamenskaya M. A. (ma_kamenskaya@mail.ru), Peoples' Friendship University of Russia, Moscow, Russia, **Khramoin I. V.** (hramoin@isa.ru), **Smirnov I. V.** (ivs@isa.ru), Institute for Systems Analysis of RAS, Moscow, Russia

The paper considers two data-driven methods for anaphora resolution of Russian texts. These methods are based on machine learning with annotated corpora and using no additional information except linguistic features. The first method uses Support Vector Machine as learning and classifying algorithms, the second method uses Decision Tree inducer. We evaluate the performance of the methods with several feature sets and corpora. Feature sets included morphological, syntactic and semantic features. In this paper we also evaluate how semantic features,

namely semantic roles, impact the performance of anaphora resolution in Russian. We used our manually annotated corpus as well as a corpus provided by the organizing committee of the forum for the evaluation of linguistic text analysis systems, an event of Dialogue 2014. Experiments showed that precision of SVM is higher on experimental data for almost all cases. It was shown that semantic features enhance the performance of the methods for anaphora resolution of Russian texts. We have also calculated the optimal distance between the anaphor and the hypothetic antecedent and used it in our methods.

PRAGMATIC ASPECTS OF INTERNET COMMUNICATION: TOWARDS WEBSITES GENRE MODELS

Kononenko I. S. (irina_k@cn.ru), A.P. Ershov Institute of Informatics Systems, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

A two-level multifaceted genre classification is proposed to cover pragmatic aspects of communication on the Web. Genre categories of websites and genre types of site constituents (pages and structural blocks) are represented as vectors of relevant pragmatic features. Praxeological parameters (activity subject, beneficiary, product, environment) are involved to represent human activity that underlies communication and manifests itself in the site structure, content and form of site constituents. Communicative parameters encompass the hierarchy of communicative tasks (including anticipated reactions of the target audience), functionality of site constituents, and the affordances of communication channel (interactivity, multimodality, and dynamics of content). Functions of site constituents together with medium features are exemplified to determine genre types of pages. The type of a textual page corresponds to a certain genre schematic structure composed of content blocks. The extraction of genre schemata is possible using the so called genre markers (cue words and constructions) that are formalized as lexico-grammatical patterns provided with format conditions.

PRACTICAL ASPECTS OF LONG-TERM ONTOLOGY-BASED INFORMATION EXTRACTION

Kravchenko A. (anna.kravchenko@interfax.ru), **Pivovarov V.** (vasiliy.pivovarov@interfax.ru), **Zharikov A.** (alexander.zharikov@interfax.ru), Interfax, Moscow, Russia

'Ontology-based information extraction' is a subfield of information extraction, where ontologies play an essential role in the process, shaping both system input and target output. There are many different approaches to creating and maintaining an ontology and little work has been done to evaluate and compare the effectiveness of those approaches.

In addition, the practical applications of those systems differ drastically from theory. Architecture that shows good performance in a single test does not necessarily perform as well in the long term. We conducted an experiment to explore the issues that arise during practical application of OBIE methods and to describe the behavior of ontologies maintained during a long period of time.

In this article we discuss emerging problems and propose working solutions for them as well as the way of evaluation of OBIE systems. Those solutions were successfully implemented in the scan-interfax.ru project and have provided sufficient quality for the commercial use of an advanced entity-based search engine extracting information from news.

HUMAN BODY IN A DIALOG: THE ORIENTATION OF SOMATIC OBJECTS IN ITS CONNECTION WITH HUMAN RELATIONS

Kreydlin G. E. (gekr@iitp.ru), **Pereverzeva S. I.** (P_Sveta@hotmail.com), Russian State University for the Humanities, Moscow, Russia

The main objective of the paper is to examine relations between corporeal, or somatic objects and some psychological aspects of human behavior, namely the relations between the communicators in a dialog. Somatic objects have been investigated from different points of view. Mostly, linguists and specialists in nonverbal semiotics have described names, features and significant actions performed by or with different somatic objects, virtually leaving aside sign manifestations of correlations between physical (corporeal) and psychological (ethical, aesthetical, etc.) aspects of human behavior. It is well-known that if a man is lengthy or extremely short or if he

is too fat or scrawny he feels bad about his deficiency. Also, it is known that many corporal defects impede or aggravate proper communication. Here we undertake a few preliminary steps in solving the problem of the systematic description of the correlations between physical and psychological properties of humans. We consider one corporal feature — "spatial orientation" — that many body parts possess and describe its relations with the psychological characteristics of interlocutors. The explication of the notion "orientation of somatic object" is given and two Russian linguistic representations of spatial orientation are discussed. The linear representation corresponds to the linguistic construction *X V Y-Instr Prep Z,* where X is an oriented object, V is a verb of orientation (e.g. *smotret' na chto-libo* 'to face smth. (e.g. about the buildings)', *byt' napravlennym na chto-libo* 'to be directed at smth.'), Y is a <so called> salient part of the object X, Prep is a preposition and Z is an orienting point. The angular representation accords with the construction *X V Prep Z pod uglom* 'at an angle' *Q* (where Q is a degree of the angle).
The basic part of the paper is devoted to the correspondence between the corporal orientations which are computed by these representations and which are expressed either verbally or nonverbally (based on the Russian body language) and ethical features of humans participating in an actual dialog. Thus, different types of bows conform regularly to the features 'respect to the addressee', 'veneration of the addressee' or just 'warm feeling to him / her'.

A DATABASE OF RUSSIAN VERBAL FORMS AND THEIR FRENCH TRANSLATION EQUIVALENTS

Kruzhkov M. G. (magnit75@yandex.ru), IPI RAS, Moscow, Russia; **Buntman N. V.** (nabunt@hotmail.com), MSU, Moscow, Russia; **Loshchilova E. Ju.** (lena0911@mail.ru), IPI RAS, Moscow, Russia; **Sitchinava D. V.** (mitrius@gmail.com), IRL RAS, Moscow, Russia; **Zalisniak A. A.** (anna.zalizniak@gmail.com), IL RAS, IPI RAS, Moscow, Russia; **Zatsman I. M.** (izatsman@yandex.ru), IPI RAS, Moscow, Russia

The paper presents the results of a project aimed at the development of methodology and information technology for the creation of a corpus-based linguistic database of verbal forms with their translation equivalents (with bilingual grammatical search functions). Within the scope of the project the following results have been achieved:
1. *Methodology and information technology* for the creation of linguistic databases based on bilingual parallel corpora have been developed (including corpora with multiple translation variants).
2. The *polyvariant parallel subcorpus* which includes Russian literary works with French translations has been created within the Russian National Corpus (RNC). Some of the parallel texts in the subcorpus include multiple translation variants.
3. On the basis of the polyvariant Russian-French corpus a *database of Russian verbal lexico-grammatical forms* (LGFs) and their French translation equivalents has been created. Equipped with bilingual grammatical search functions, the database is a unique resource that can be used for investigating a wide range of various cross-linguistic problems.
4. A number of concepts in the areas of Russian verbal categories and Russian-French contrastive grammar have been refined.

CONDITIONAL RANDOM FIELD IN SEGMENTATION AND NOUN PHRASE IINCLINATION TASKS FOR RUSSIAN

Kudinov M. S. (m.kudinov@samsung.com), Samsung R&D Institute Russia, Moscow, Russia; Dorodnitsyn Computing Center RAS, Moscow, Russia, **Romanenko A. A.** (a.romanenko@samsung.com), Samsung R&D Institute Russia, Moscow, Russia; Moscow Institute of Physics and Technology, Moscow, Russia, **Piontkovskaja I. I.** (p.irina@samsung.com), Samsung R&D Institute Russia, Moscow, Russia

We propose solutions of several NLP problems for Russian making use of the conditional random fields (CRF) framework, including: shallow parsing (chunking), temporal expressions extraction and noun phrase inflection. Each of the three problems are important in speech generation, data mining and spoken dialogs systems design. The purpose of shallow parsing is to extract from the text syntactically related word forms (e.g. noun phrases) without full parsing. It may be useful in data mining applications. Temporal expressions extraction is important for natural language understanding modules of spoken dialog systems. Usually rule-based methods are used to address this problem. Noun phrase inflection is needed for speech generation modules.

The main problem is to detect word forms for inflection. For all three problems statistical approach was taken. We use simple version of CRF named linear-chain CRF. In shallow parsing and time expressions extraction state-of-the-art results were achieved. In noun phrase inflection, the level of F_1-measure exceeded 95.

CONSTRUCTIONS WITH THE CONJUNCTION *CHTOBY*: RESOURCES AND CORRELATIONS

Kustova G. I. Kustova (galinak03@gmail.com), Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

The article deals with complex sentences with a noun in the main clause and the conjunction *chtoby* in the subordinate clause. Construction «desirable feature» (*Gde najdesh sidelku, chtoby xorosho ladila s Vasej?*) and construction «functional inconsistency» (*On ne dama, chtoby emu cvety darit'*) are compared with the resource construction (*U nas est' vremy'a, chtoby sxodit' v kino*).

THE DESCRIPTION OF LOCATIVE DEPENDENCIES IN A NATURAL LANGUAGE PROCESSING MODEL

Leontiev A. P. (Aleksey_L@abbyy.com), **Petrova M. A.** (Maria_Pet@abbyy.com), ABBYY, Moscow, Russia

The paper suggests semantic and syntactic descriptions of locative dependencies in an NLP model and focuses on the problems which locative adjuncts evoke for a system aimed at different tasks based on semantic analysis, especially at machine translation. A formal description of locative groups faces several problems. The first is the definition of locative semantic relations between words, as locative dependencies can have different meanings, such as the meanings of initial and final points (*walk [from/to the door]*), route (*walk [across the room]*), and others. Second, one has to define the set of words that can fill locative adjuncts, and the border between the locative and non-locative groups is not always distinct: *in the street* is definitely a locative, but what about *on the Internet* or *in a meeting*?

Third, the syntactic realizations of locative senses are rather numerous. On the one hand, locative adjuncts include many prepositions with different semantics—like *on, in, under, above*, etc. On the other hand, different nouns combine with different prepositions to denote the same meaning, like *in the country*, but *on the island*.

The current paper suggests a formal approach appropriate for dealing with all these difficulties.

UNIVERSAL MELODIC PORTRAITS OF INTONATION PATTERNS IN RUSSIAN SPEECH

Lobanov B. M. (Lobanov@newman.bas-net.by), **Okrut T. I.** (tatberrie@gmail.com), United Institute of Informatics Problems NAS Minsr, Belarus

We proceed from the model of intonation patterns by Elena Bryzgunova, which is widely used in the teaching of Russian speech intonation. This model includes seven patterns: IP1 (the falling tone), IP2 (the falling tone with a certain prosodic emphasis), IP3 (the rising tone with subsequent fall), IP4 (the falling-rising tone). IP5 (combination of the rising, smooth and falling tones), IP6 (combination of the rising and smooth tones), IP7 (combination of the rising tone with the glottal stop). We present a model of intonation portraits of accentual units (the PAU model), proposed by one of the authors of this paper and effectively used in the practice of Russian speech synthesis for a long time. The PAU model assumes that, for a certain intonation type, the topological properties of the melodic contour are independent of the quantitative and the qualitative characteristics of the pre-nucleus, the nucleus and the post-nucleus of accentual units. The methodology of an experiment of integration of the two models into a unified model of Universal melodic portraits of intonation patterns (UMP-IP) is discussed. The new model is shown to effectively represent the tonal structure of Elena Bryzgunova's intonation patterns and ensure the invariance of the quantitative and the qualitative constituents of the sentence pronounced as well as the pitch and the range of the speaker's voice. The obtained results are discussed from the viewpoint of applicability to the practice of teaching Russian as the second language.

RUTHES-LITE, A PUBLICLY AVAILABLE VERSION OF THESAURUS OF RUSSIAN LANGUAGE RUTHES

Loukachevitch N. V. (louk_nat@mail.ru), **Dobrov B. V.** (dobrov_bv@mail.ru), **Chetviorkin I. I.** (ilia2010@yandex.ru), Lomonosov Moscow State University, Moscow, Russia

The paper presents RuThes-lite, a publicly available version of RuThes linguistic ontology, which has been developed for more than fifteen years and is intended for automatic document processing. RuThes has considerable similarities with WordNet: 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 names of concepts, attention to multiword expressions, intentional inclusion of terms of the sociopolitical domain, a set of conceptual relations. RuThes-lite was generated from RuThes on the basis of the most frequent words in a contemporary news collection. Besides, we describe additional data, which have been specially prepared for RuThes-lite publication: morph-syntactic labeling of thesaurus text entries and assignment of glosses to concepts.

DESIGNING "HUMAN CHARACTERS" LEXICAL DATABASE

Lukashevich N. Ju. (natalukashevich@mail.ru), **Kobozeva I. M.** (kobozeva@list.ru), Lomonosov Moscow State University, Moscow, Russia

The paper discusses a general layout of "Human Characters" lexical database specifically developed to study the meanings of words from the semantic field of human character traits. It is intended as a resource providing a format for a comprehensive analysis of character words usage in different languages. A database with contexts from large modern corpora is considered a convenient tool for semantic analysis which offers such advantages as facilitating data storage and presentation, and keeping the analysis consistent while making changes possible at the same time. It is shown how several issues which significantly influence the analysis procedures are resolved in the pilot database version. These include identifying relevant contexts, describing features of a typical situation in which the character trait in question is exhibited, and comparing contextual meanings of the studied words. The suggested technique provides a more flexible tool for capturing similarities and differences between contexts within one language on the one hand, and gives ground for comparing the usage of translation equivalents on the other.

EVALUATION OF FRAME-SEMANTIC ROLE LABELING IN A CASE-MARKING LANGUAGE

Lyashevskaya O. N. (olesar@gmail.com), National Research University Higher School of Economics, Moscow, Russia; Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia, **Kashkin E. V.** (egorkashkin@rambler.ru), Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

The paper discusses evaluation techniques for semantic role labeling in Russian. It has been shown that the quality of FrameNet-style semantic role labeling largely depends on the quantity of roles and may decrease if the inventory of roles in the training set differs from that in the output resource. Our study is the first step towards the 'smart' evaluation tool which would introduce linguistically relevant criteria to evaluation; be able to put the mistakes on a scale from minor to critical ones; make evaluation easier in case the grid of roles varies.

We run an experiment based on the data from the Russian FrameBank, a FrameNet-oriented open access database which includes a dictionary of Russian lexical constructions and a corpus of tagged examples. The semantic role is one of the parameters that define the predicate-argument patterns in FrameBank. The inventory of roles is modeled hierarchically and forms a graph. We explore the cases when the role induced by the system and the answer of the gold standard do not match. We analyze the statistical criteria of distribution of roles in the patterns and the distance between the source and the target in the graph of roles as a mean to assess the goodness of fit.

INTERNET DATA IN THE STUDY OF LANGUAGE CHANGE: A CASE STUDY OF ALTERNATIONS IN RUSSIAN COMPARATIVES AND A PROGRAM TO WORK WITH SUCH DATA

Magomedova V. D. (varya.magomedova@gmail.com), Saint Petersburg State University, St. Petersburg, Russia, **Slioussar N. A.** (slioussar@gmail.com), National Research University Higher School of Economics, Moscow, Russia Saint Petersburg State University, St. Petersburg, Russia

The Internet is a unique source of non-standard forms, which gives us a novel opportunity to analyze fine-grained dynamics of language change. We used this opportunity to study the decay of historic consonant alternations in Russian. In standard Russian, these alternations are present in some verb forms and in comparatives (e.g. *suxoj* 'dry' — *sushe* 'drier', *ljubit'* 'to love' — *ljublju* 'I love'), as well as before certain derivational suffixes. Verb forms have been recently studied by Slioussar and Kholodilova (2013), and we looked at comparatives. Two groups of adjectives were selected: ones that have normative comparatives with alternations and ones that do not, but native speakers still try to generate such forms. In the first group, some adjectives like *ubogij* 'poky' have up to 30% of comparatives without alternations, but, unlike with verbs, no significant correlation with adjective frequency or its other characteristics was found. The second group consisted primarily of compound adjectives ending in *-gij, -kij, -xij*. Here, the most important factor is whether the second part of the compound is used as an independent adjective. If it is not (e.g. as in *dlinnorukij* 'long-armed'), most comparatives lack alternations. Searching for forms on the Internet, we faced many problems. The counts provided by search engines are extremely inaccurate, only the first thousand results are shown, they cannot be downloaded in a convenient format, contain a lot of typos and other irrelevant data etc. We present a program called *Lingui-Pingui* that we developed to solve these and some other problems.

A MULTI-FACETED APPROACH TO REFERENCE RESOLUTION IN ENGLISH AND RUSSIAN

McShane M. (mcsham2@rpi.edu), Rensselaer Polytechnic Institute, Troy, NY, USA

This paper argues that the detection and resolution of referring expressions can be profitably distributed across modules of a language processing system, rather than being bunched at the end of a text analysis pipeline. The approach is being implemented within the OntoAgent cognitive architecture, which supports the development of multi-functional, language-endowed agents that can collaborate with people in task-oriented applications. Although current development within OntoAgent orients around English, the architecture itself and most of its knowledge bases are language-independent. Drawing upon my past descriptive work on reference and ellipsis in Russian, I will suggest how the same reference resolution strategies might be applied to this and other languages. More generally, I will motivate the need to approach linguistic phenomena in a holistic paradigm, rather than as highly compartmentalized subtasks, which has become the norm for natural language processing applications.

DUSHI SIRENEVAJA CVET'... OR JUST A NONSENSE (*KAKAJA-TO KHREN'*)? NOUNS WITHOUT SUFFIXS IN THE TEXTS OF RUSSIAN AUTHORS

Mikheev M. Ju. (m-miheev@rambler.ru), Research Computing Center of Lomonosov Moscow State University, Moscow, Russia

Nouns without suffixes, feminine, the 3rd declination *(e. g. ludskaya molv')* and masculine, the 2nd declination (e. g. *konski top)* were honored by Alexander Pushkin as legitimate *root* Russian words. His friend, Vladimir Dal managed to «expand» his dictionary precisely thanks to these words. At the turn of the XIX–XX century these words, especially the first group, became very frequent in Russian poetry and prose. Some of them were recreated. We can find many interesting examples in Sergei Yesenin's and Mikhail Sholokhov's texts. The latter author, made out of dialect and colloquial words distinct markers of his style.

TOWARDS A FINE-GRAINED DESCRIPTION OF INTENSIFYING ADJECTIVES FOR TEXT PROCESSING

Milichevich J. (jmilicev@dal.ca), Dalhousie University, Halifax, Canada,
Timoshenko S. (timoshenko@iitp.ru), Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia

We address collocations of the type "*Intensifying Adjective + NOUN*", such as *heavy RAIN* and *complete DISAGREEMENT*, known as Magn type collocations. Such a collocation can be represented as a functional dependency: Magn(*RAIN*) = *heavy*, where Magn is a (lexical) function responsible for the meaning 'very'/'high degree', and *heavy* the value that Magn has with *RAIN*, its keyword. The formalism of lexical functions has proved its usefulness in various NLP tasks, but on close inspection its semantic granularity turns to be insufficient. We propose a refinement of the notion of Magn by distinguishing Magn's semantic subtypes. Our description, which proceeds from the assumption that a choice of a Magn type collocate is not arbitrary, takes into account the following factors:

- semantic class of the keyword (= its semantic label, corresponding to the generic semantic component of its definition) and/or its actants;
- semantic component(s) in the keyword's definition targeted by intensification;
- semantic contrasts observed among Magn type collocates of a given keyword.

We tested our approach on data from the Russian and English explanatory-combinatorial dictionaries developed for the multi-purpose language processing system ETAP-3. As our results show, Magn's semantic subtypes we have identified allow for the encoding of lexicograpahic information in a way that is not only precise but also has predictive power.

NEOLOGISMS ON FACEBOOK

Muravyev N. A. (nikita.muraviev@gmail.com), Digital Society Laboratory LLC, Moscow, Russia; Moscow State University, Faculty of Theoretical and Applied Linguistics, Moscow, Russia, **Panchenko A. I.** (a.panchenko@digsolab.com), Digital Society Laboratory LLC, Moscow, Russia; Universite catholique de Louvain, Louvain-la-Neuve, Belgium,
Obiedkov S. A. (sergei.obj@digsolab.com), Digital Society Laboratory LLC, Moscow, Russia; National Research University Higher School of Economics, Department of Applied Mathematics and Information Science, Moscow, Russia

In this paper, we present a study of neologisms and loan words frequently occurring in Facebook user posts. We have collected a dataset of over 573 million posts written during 2006–2013 by Russian-speaking Facebook users. From these, we have built a vocabulary of most frequent lemmatized words missing from the Opencorpora dictionary (http://opencorpora.org/dict.php) the assumption being that many such words have entered common use only recently. This assumption is certainly not true for all the words extracted in this way; for that reason, we manually filtered the automatically obtained list in order to exclude non-Russian or incorrectly lemmatized words, as well as words recorded by other dictionaries or those occurring in pre-2000 texts from the Russian National Corpus (http://www.ruscorpora.ru). The result is a list of 168 words that can potentially be considered neologisms.

We present an attempt at an etymological classification of these neologisms (unsurprisingly, most of them have recently been borrowed from English, but there are also quite a few new words composed of previously borrowed stems) and identify various derivational patterns. We also classify words into several large thematic areas, "internet", "marketing", and "multimedia" being among those with the largest number of words.

We consider our results preliminary, but believe that, together with the word base collected in the process, they can serve as a starting point in further studies of neologisms and lexical processes that lead to their acceptance into the mainstream language.

CONDITIONAL RANDOM FIELD FOR MORPHOLOGICAL DISAMBIGUATION IN RUSSIAN

Muzychka S. A. (s.muzychka@samsung.com), Lomonosov Moscow State University, Moscow, Russia; Samsung R&D Institute Rus, Moscow, Russia, **Romanenko A. A.** (a.romanenko@samsung.com), Moscow Institute of Physics and Technology, Moscow, Russia; Samsung R&D Institute Rus, Moscow, Russia, **Piontkovskaja I. I.** (p.irina@samsung.com), Samsung R&D Institute Rus, Moscow, Russia

We consider the problem of morphological disambiguation in Russian using statistical methods; specifically, we apply conditional random field (CRF). We propose a new modified model of linear chain CRF, which demonstrates results close to the state-of-the-art. We also propose a new

statistical approach to text normalization problem using CRF. Namely, we solve the problem of normalization of numerals written as digits. Our approach allows for the consideration of both cardinal and ordinal numbers.
In order to train and test our models we used Russian text corpora. For morphological disambiguation, we used data from OpenCorpora and the SynTagRus linguistic corpus. For number normalization we used the Russian National Corpora (RusCorpora).
A brief overview of the CRF model is given, followed by a detailed description of the applied algorithm, assumptions on the training and test set, and a description of features for each particular issue.

"VCHERA NASOCHINYALSYA VOROH STROK": PRODUCTIVE CIRCUMFIXAL INTENSIFYING PATTERNS IN RUSSIAN

Nedoluzhko A. Yu. (nedoluzko@ufal.mff.cuni.cz), Charles University in Prague, Prague, Czech Republic, **Khoroshkina A. S.** (annakhor@gmail.com), Lomonosov Moscow State University, Moscow, Russia

The current paper addresses verbal circumfixal derivation patterns in modern Russian. The discussion is focused on a series of circumfixes which trigger the intensified usage of the basic verb (~'*keep doing P too much*'). Derivatives built up by adding a prefix and a reflexive *-ся* to an imperfective verb are examined. Although each prefix adds specific shades of meaning to the verb, such patterns are, however, claimed to share common features at different levels of linguistic analysis, such as morphology, syntax, and semantics. Furthermore, such patterns are highly productive in modern language; once certain constraints are fulfilled, an intensified derivative can be formed from any imperfective verb. This fact, along with the patterns in question sharing certain common features, allow us to argue that they can be considered inflectional, rather than derivational.

A SUMMARIZATION MODEL BASED ON THE COMBINATION OF EXTRACTION AND ABSTRACTION

Osminin P. G. (osperevod@gmail.com), South Ural State University, Chelyabinsk, Russia

We suggest a model of automatic summarization for scientific and technical texts. This model combines extractive and abstractive approaches for summarization and was developed on the basis of comparative analysis of authors' summaries and full texts of corresponding papers. The model consists of three main components: a keyword extractor, a domain and task oriented static knowledge base and a summarization algorithm. The keyword extractor is off-the shelf tool LanAKey_Ru, adapted to the application. Static knowledge includes stop lexicons, conceptual net, templates for summary content selection and rules for the generation. Stop lexicons are used for removing text segments irrelevant for the document summary. The conceptual net is used for semantic analysis of a document text helping content selection. Templates for information extraction are frame structures. Their slots are to be filled with extracted fragments of document sentences. Rules for summary generation define the grammar of summary sentences and their order. The summarization algorithm consists of four top level procedures—preprocessing, analysis, content selection and summary text generation. The model is described on the example of Russian scientific papers in mathematical modeling domain.

SUSPENDED ASSERTION AND NONVERIDICALITY

Paducheva E. V. (elena.paducheva@yandex.ru), VINITI RAN, Moscow, Russia

Two notions are compared: *suspended assertion* and *nonveridicality*. It is argued that these notions, though used in the frameworks of different linguistic theories, are applied to similar linguistic phenomena. In this paper the notion of nonveridicality is applied to one group of Russian indefinite pronouns — namely, to **negative polarity pronouns** (NPP). Four groups of non-referential indefinite pronouns are differentiated in Russian: negative pronouns (*ni-* series), non-specific indefinite (*-nibud'* series), free choice (*ugodno* series and *ljuboj*) and negative polarity pronouns (*-libo* and *by to ni bylo* series). Following Giannakidou 1998, I reject the hypothesis that NPPs are licensed in the context of downward entailment operators only. I also argue against what is claimed in Giannakidou 2011, that NPPs are licensed in the three types of environments: negative, downward entailing and nonveridical: all contexts of the Russian NPPs can be demonstrated to be nonveridical, and the context of negation is one of them. The list of contexts licensing all the four classes of non-referential pronouns is suggested. Each of the four classes of pronouns chooses its own subset of contexts from the list.

SENTIMENT INDEX OF THE RUSSIAN SPEAKING FACEBOOK

Panchenko A. I. (alexander.panchenko@uclouvain.be), Universite catholique de Louvain, Louvain-la-Neuve, Belgium; Digital Society Laboratory LLC, Moscow, Russia

A sentiment index measures the average emotional level in a corpus. We introduce four such indexes and use them to gauge average "positiveness" of a population during some period based on posts in a social network. This article for the first time presents a text-, rather than word-based sentiment index. Furthermore, this study presents the first large-scale study of the sentiment index of the Russian-speaking Facebook. Our results are consistent with the prior experiments for English language.

GOVERNMENT OF THE BORROWED NEOLOGISMS DENOTING OBJECTS OF FILM INDUSTRY

Pestova A. R. (pestova2012@gmail.com), Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

The present paper deals with the government of borrowed neologisms, denoting objects of film industry: *трейлер* 'trailer', *тизер* 'teaser', *ремейк* 'remake', *сиквел* 'sequel', *приквел* 'prequel', *триквел* 'trequel', *квадриквел* 'quadriquel', *мидквел* 'midquel' and *интерквел* 'interquel'. Dictionaries don't give any information about syntactical features of these words. The study shows that government of these nouns is variational and the revealed constructions are synonymous and redundant. As language tends to eliminate redundancy, we tried to find the most popular variant for each word. The Statistics of Internet resources "Yandex.Novosti" (news segment), "Yandex.Blogi" (blogosphere) and corpus RuTenTen was analysed. All listed nouns tend to govern non-prepositional genitive. The used method can be applied to other borrowed neologisms, for example to nouns, referring to music scene (*ремикс* 'remix', *кавер* 'cover version', *(видео) клип* 'video clip' and *(видео)ролик* 'video clip'). They prefer prepositional-case construction *на* + accusative.

PRAGMATICS OF STRIKETHROUGH: NORMS OF COMMUNICATION AND OPTIMALITY THEORY

Piperski A. Ch. (apiperski@gmail.com), **Somin A. A.** (somin@tut.by), Russian State University for the Humanities, Moscow, Russia

The paper presents a description of intentional strikethrough on the Web using a combination of theories from pragmatics and phonology, namely the theory of implicature by Grice (1975), the politeness theory by Brown and Levinson (1978, 1987), and Optimality Theory. We argue that the study of this phenomenon can shed light on some more general aspects of communication theory, such as the mechanism of choosing one viewpoint among many options. We also describe graphical and verbal substitutes for strikethrough in blogs and literary works.

"THEY SHOT HIM DEAD, OH, NO, THEY KNIFED HIM DEAD WITH A SABER": SELF-REPAIRS IN ORAL STORIES

Podlesskaya V. I. (podlesskaya@ocrus.ru), Russian State University for the Humanities, Moscow, Russia

The paper introduces a discourse oriented classification of repair types in Russian by addressing, inter alia, the following questions: (i) whether or not self-repairing entails speech disfluency; (ii) whether or not the fragment under repair and its repaired correlate are structurally isomorphic; (iii) does the speaker revise a lexical, a morpho-syntactic, or a phonologic shape of the reparandum. Basing on the data from the Prosodically Annotated Corpus of Spoken Russian, established classes of repairs were analyzed qualitatively and quantitatively. Fluent isomorphic repairs appeared to be the most frequent in the corpus, although fluent non-isomorphic repairs, as well as disfluent isomorphic and disfluent non-isomorphic repairs are also attested.

ANAPHORIC ANNOTATION AND CORPUS-BASED ANAPHORA RESOLUTION: AN EXPERIMENT

Protopopova E. V. (protoev@gmail.com), **Bodrova A. A.** (anastasie.bodrova@gmail.com), **Volskaya S. A.** (svetlana.volskaya@gmail.com), **Krylova I. V.** (krylova93@gmail.com), **Chuchunkov A. S.** (scarywound@gmail.com), Saint Petersburg State University, St. Petersburg, Russia, **Alexeeva S. V.** (alexeeva@opencorpora.org), **Bocharov V. V.** (bocharov@opencorpora.org), **Granovsky D. V.** (granovsky@opencorpora.org), OpenCorpora.org

The paper describes the noun phase and anaphora annotation in OpenCorpora and compares it to that in other corpora. We discuss the choice of representative texts for anaphoric annotation and the basic principles of syntactic annotation. In case of noun phrase annotation we followed the scheme introduced earlier for morphological annotation: it was carried out in two stages: firstly, all noun phrases and some other syntactic units were annotated by a heterogenous group of people, then a linguist compared all markup results and found the best one, or corrected mistakes. We present some annotation results and cases of annotator's disagreement and proceed to introduce our data-driven anaphora resolution system based on decision trees. We then list the features used to fit the classificator and discuss their relevance and some changes which improved the classificator performance. We also present out rule-based approach to automated noun phrase extraction using Tomita parser. A baseline for anaphora resolution is introduced and we compare it with our results.

RECENT ADVANCES IN (DEEP) REPRESENTATION LEARNING

Schütze H. (hs0711@cis.uni-muenchen.de), University of Munich, Germany

Traditionally, natural language processing (NLP) systems have made use of resources compiled by (computational) linguists based on linguistic theory that provide rich information about linguistic objects. For example, computational lexica specify morphological paradigms and subcategorization frames of verbs. In contrast, statistical NLP systems frequently start out with no explicit representation of linguistic objects and instead learn what they need from training data on a task-by-task basis. A third approach—which has gained much interest recently—is to learn generic representations of linguistic objects and then reuse them for a wide variety of tasks. Its premise is that giving an NLP system non-task-specific generic information about words and other linguistic objects will help it in performing well at a particular task.

Examples of such generic representation models include the vector space model, dimensionality reduction, clustering and deep learning. I will review recent research results in representation learning and discuss benefits and drawbacks of the three approaches.

ON THE CLASS OF RUSSIAN PARAMETRIC ADVERBS

Semenova S. Ju. (sonya_sem@mail.ru), INION RAS, Russian State University for the Humanities, Moscow, Russia

The paper deals with Russian parametric adverbs i. e. those revealing the values of the quantitative parameters (*gluboko* [deeply], *chasto* [often / thickly / frequently], *redko* [rarely / seldom], *bystro* [rapidly / quickly], *izdaleka* [from afar] etc). Characteristics of parametric adverbs seem to be much less investigated (in particular, in the perspective of information extraction) than those of parametric nouns, adjectives, and verbs. A number of grammatical and semantic groups of adverbs are presented. The parametric meaning is found to be distributed among various traditional grammatical and semantic classes of adverbs. For parametric adverbs morphologically derived from adjectives, we discuss semantic priority or lack thereof with respect to adjectives. The parametric meaning can take place for a secondary sense of an adverb, so that ambiguity, connotations, and implication are essential in the descriptions aimed at information extraction. The correspondence between the quantitative meaning of the adverb and the name of the physical value (*izdaleka — rasstojanie* [distance]) are considered. Corpora examples of various types of parametric data coded with the help of parametric adverbs are presented.

DISTRIBUTIONAL-STATISTICAL ANALYSIS OF REGIONAL PRESS (NEWSPAPERS OF GRODNO REGION OF BELARUS)

Shaikevich A. Y. (lingstat@yandex.ru), **Savchuk S. O.** (savsvetlana@mail.ru), Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

The paper is an application of distributional-statistical analysis (DSA) to the sub-corpora of Grodno region newspapers corpus. The sub-corpora under study are district newspapers, "The Evening Grodno" and commentaries to the latter. With the help of DSA hundreds of keywords have been elicited for each sub-corpus. The linguistic interpretation of those three lists showed that the keywords grouped into clusters reflect both thematic and stylistic features of the sub-corpora.
The district newspapers are specific in the choice of domains (mostly of local interest) and stylistic flavor (mostly official and bookish, to some extent resembling Soviet use). "The Evening Grodno" is more colloquial stylistically; its domains are naturally connected with the day-to-day city life and some topics which were unexpected, such as a large cluster of words denoting places of interest for tourists and inhabitants of the city. The keywords of the commentaries brings the stylistic trend of "The Evening Grodno" to its logical end. The method may be used for comparative analysis of other corpora, which might bring about new results depending on the composition of the corpus.

PERLOCUTIONARY SPEECH ACTIONS AND PERLOCUTIONARY VERBS

Shatunovskiy I. B. (shatun49@mail.ru), International University for Nature, Society, and Man "Dubna", Dubna, Russia

Perlocutionary verbs like *ubezhhdat'* 'to convince / persuade', *nastaivat'* 'to insist', *ugovarivat'* ≈'to persuade', *uspokaivat'* 'to calm', *objasn'at'* 'to explain', *xvastatsy'a* 'to boast' etc. are verbs denoting perlocutionary actions. Perlocutionary actions, as defined in the paper, are unconventional actions performed by means of conventional illocutionary acts. Perlocutionary actions are aimed to achieve certain effects, goals, but they do not necessarily achieve them. Perlocutionary verbs such as *preduprezhdat'* (*to warn*), *nastaivat'* (*to insist*), *uveryat'* ('to assure') can turn into illocutionary verbs. In this case the perlocutionary text is contracted and some parts of it are taken in the meaning of the verb becoming a sign of that contraction. Perlocutionary actions and verbs can be divided into several groups according to supposed goals and effects of a perlocutionary action. They are: (1) perlocutionary actions having a clear aim which is embedded, fixed in the meaning of the verb denoting that action; this aim can be achieved or not; (2) perlocutionary actions that do not have a clear aim, but have a bundle of possible aims that are not fixed in the meanings of the corresponding perlocutionary verbs; (3) perlocutionary (and some illocutionary) actions that have a clear aim, and that aim is achieved any time the speaker does that action. These groups differ with respect to the meaning of their perfective forms. In the paper these differences are described and explanations for semantic peculiarities of the perfective forms are proposed.

METHODS FOR SEMANTIC ROLE LABELING OF RUSSIAN TEXTS

Shelmanov A. O. (shelmanov@isa.ru), **Smirnov I. V.** (ivs@isa.ru), Institute for Systems Analysis of the Russian Academy of Sciences, Moscow, Russia

The paper introduces two methods for semantic role labeling of Russian texts. The first method is based on semantic dictionary that contains information about predicates, roles and syntaxeme features that correspond to the roles. It also uses heuristics and integer linear programming to find the best joint assignment of roles. The second method is data-driven semantic-syntactic parsing, which was implemented using MaltParser. It performs transition-based data-driven parsing simultaneously building a syntactic tree and assigning semantic roles. It was trained with various feature sets on SynTagRus Treebank, which was automatically enriched with semantic roles by the dictionary-based parser. We managed to automatically alleviate mistakes in the training corpus using output of the data-driven parser. We evaluated the performance of the parsers on the subcorpus of SynTagRus, which we manually annotated with semantic information. The dictionary-based parser and the data-driven semantic-syntactic parser showed close performance. Although the data-driven parser did not outperform the dictionary-based parser, we expect that it can be beneficial in some cases and has potentials for further improvement.

THE DIALECTAL SUBCORPUS WITHIN THE RUSSIAN NATIONAL CORPUS: TODAY AND TOMORROW

Sitchinava D. V. (mitrius@gmail.com), Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia, **Kachinskaya I. B.** (kacza@yandex.ru), Lomonosov Moscow State University, Moscow, Russia

The main results of the project aimed at developing the dialectal subcorpus of the RNC were the creation of a pilot corpus and the change of the markup principles encompassing many dialectoloigical parameters. A working place program was created and many texts were marked up using the new technology. The present goal of our team is a considerable increase of the corpus, its representativeness and the depth of linguistic processing. The dialectal texts available for search in the RNC (www.ruscorpora.ru/search-dialect.html) will be considerably updated, with the overall corpus size reaching 1 mln tokens. The texts, mainly unpublished or published in rather obscure editions, are to be made available for a wider circle of dialectologists. Some texts are to be accompanied with video and audio. Alongside with word-by-word grammarical markup with resolved homonymy, the texts are to be tagged extensively on the metalevel (data of creation, dialect, overall phonetical properties and others). The accumulation of dialectal texts will be continued, the dialectologists who had collected valuable texts are invited to share their results with the professional community.

USING LATENT SEMANTIC ANALYSIS FOR SIMULATING OF CHILDREN'S COGNITIVE DEVELOPMENT

Solovyev A. (lechat1@mail.ru), St. Petersburg State University, St. Petersburg, Russia

In the 20th century Noam Chomsky formulated the so-called Plato's problem: why is the amount of our knowledge much greater than we can extract from our everyday experience? For example, the vocabulary of preschool children (aged 6–7) averagely increases by 3–8 words every day, and not every word refers to any reality or action (for example, abstract concepts, words carrying "phatic" or uninformative assignment, etc.).

How does the child recognize each new meaning of the word and its relation to others, or why are new "meanings" formed? We propose a method to simulate associative-semantic relations between words. On the one hand. it eliminates rigid binding of a lexical unit to any cluster, and on the other it presents a complete system of relationships between words.

The paper presents the results of three experiments with cognitive development of 4–7-year-old children using a Latent Semantic Analysis (LSA) that permits comparisons of semantic similarity between pieces of textual information. We used a technique developed by G. Denhière and B. Lemaire. The principal distinctions of our research are that for the first time, the experiments were performed 1) on the Russian language; and on pre-school children. The children were grouped into two categories: 4–5 and 6–7 years, which corresponds to age variability of cognitive development.

Two experiments describe semantic and associative similarity between LSA models and the children's cognitive development. The third experiment describes using LSA to measure the children's semantic memory. The results are compared to children's model data and adults' model data. The computational models are built from the LSA of a multisource child corpus and of an internet mass media corpus.

Our findings confirm that: 1) LSA can be used to simulate a variety of children's cognitive processes; 2) LSA models represent the development of different age groups children's cognitive processes, in particular associative semantic processes and short-term and long-term memory work; 3) this method may be recommended for the comparative study of children's cognitive development, in particular, the development of associative-logical thinking, verbal discourse, the development of memory.

ASSOCIATING SYMPTOMS WITH SYNDROMES. RELIABLE GENRE ANNOTATION FOR A LARGE RUSSIAN WEBCORPUS

Sorokin A. (alexey.sorokin@list.ru), **Katinskaya A.** (a.katinsky@gmail.com), **Sharoff S.** (s.sharoff@leeds.ac.uk), Lomonosov Moscow State University, Russian State University for the Humanities, Moscow, Russia; University of Leeds, Leeds, UK

The paper describes several experiments aimed at establishing the parameters for genre annotation of potentially any text which can be collected from the Russian web. We started with a set of text classification parameters, refined them iteratively in several studies and established a reliable framework, which was further subjected to clustering analysis. Overall, we obtained the level of agreement for Krippendorff's α to be in the range of $0.51<\alpha<0.84$. We have also discovered the most common combinations of parameters in the test corpus, which should form the basis for classifying very large samples of the Russian web.

A PRODUCTION SYSTEM FOR INFORMATION EXTRACTION BASED ON COMPLETE SYNTACTIC-SEMANTIC ANALYSIS

Starostin A. S. (astarostin@abbyy.com), **Smurov I. M.** (ismurov@abbyy.com), **Stepanova M. E.** (mstepanova@abbyy.com), ABBYY, Moscow, Russia

The article presents a mechanism for information extraction from unstructured natural language data. The key feature of this mechanism is that it relies on deep syntactic and semantic analysis of the text. The system takes a collection of syntactic-semantic dependency trees as input and, after processing them, outputs an RDF graph consistent with certain domain ontology.

The mechanism was implemented within a deployable information extraction system, which is a part of ABBYY Compreno technology—a powerful tool for a broad range of NLP-tasks that include machine translation, semantic search and text categorization. The description of the extraction algorithm and the results of the system performance evaluation are given.

Evaluation tests were conducted on the MUC-6 corpus. The overall F-measure we achieved using Compreno technology was 0.83, which is lower than the best results claimed by the researchers using machine learning approaches. Our system is still under development at the moment and we hope to improve its performance in the future. One of the advantages of Compreno technology is that, unlike many statistical approaches, it does not show an abrupt performance drop if the test corpus is changed. Thus Compreno demonstrates little dependence on the exact textual data it receives and therefore might be seen as a more universal and less domain-dependent solution. Our tests on the CoNLL corpus yielded an F-measure of 0.75 with no prior adjustments made.

THE EXPERIENCE OF BUILDING INDUSTRIAL-STRENGTH PARSER FOR ARABIC

Strebkov D. Y. (strebkov@dictum.ru), **Hilal N. R.** (hilal@dictum.ru), **Redjaimia A.** (redjaimia@dictum.ru), **Skatov D. S.** (ds@dictum.ru), Dictum Ltd., Nizhny Novgorod, Russia

We present a propagation of a hybrid approach for natural language parsing on Semitic languages on the example of the Arabic language. The hybrid approach proposes a way for acquiring dependency and constituency parses simultaneously at every step of the analysis. The result of the propagation is represented by a syntactic parser for Arabic language and the fact that the parser shows quite satisfactory results and belongs to the group of rule-based parsers actually forms scientific novelty of this article. We give a short review of Arabic Natural Language Processing (NLP) technologies and their current state and then describe steps that were required for our propagation: choosing of morphological analyzer, morphological index compression scheme, description of rule base system that is used by the parser, modifications that were needed for tuning in the core parsing algorithm. We also designate problems that we faced during the propagation and the results that we finally achieved. In the end we provide results of brief evaluation of the parser and give information on its current usage.

RU-EVAL-2014: EVALUATING ANAPHORA AND COREFERENCE RESOLUTION FOR RUSSIAN

Toldova S. Ju.[1,2] (toldova@yandex.ru), **Roytberg A.**[1,3] (cvi@yandex.ru), **Ladygina A. A.**[2] (aladygina@yahoo.com), **Vasilyeva M. D.**[2] (linellea@yandex.ru), **Azerkovich I. L.**[2] (iazerkovich@gmail.com), **Kurzukov M.**[2] (mkurg@ya.ru), **Sim G.**[2] (sim.ge@yandex.ru), **Gorshkov D. V.**[2] (d.gorshkoff@gmail.com), **Ivanova A.**[2] (ivanastas@gmail.com), **Nedoluzhko A.**[4] (nedoluzhka@gmail.com), **Grishina Y.**[5] (jul_gr@mail.ru),
[1] Natioanl Research University Higher School of Economics, Faculty of Philology, Myasnitskaya 20, 101000 Moscow, Russia, [2] Moscow State University, Philological Faculty, Dept. of Theoretical and Applied Linguistics, Leninskie gory, GSP-1, 119991 Moscow, Russia, [3] IMPB (Institut of mathematical problem in biology) RSS, [4] Charles University in Prague, [5] Applied Computational Linguistics, University of Potsdam

The paper reports on the recent forum RU-EVAL—a new initiative for evaluation of Russian NLP resources, methods and toolkits. The first two events were devoted to morphological and syntactic parsing correspondingly. The third event was devoted to anaphora and coreference resolution. Seven participating IT companies and academic institutions submitted their results for the anaphora resolution task and three of them presented the results of the coreference resolution task as well. The event was organized in order to estimate the state of the art for this NLP task in Russian and to compare various methods and principles implemented for Russian. We discuss the evaluation procedure. The anaphora and coreference tasks are specified in the present work. The phenomena taken into consideration are described. We also give a brief outlook of similar evaluation events whose experience we lay upon. In our work we formulate the training and Gold Standard corpora construction guidelines and present the measures used in evaluation.

ON DERIVED PREPOSITIONS: ADVERBAL PREPOSITIONS

Uryson E. V. (uryson@gmail.com), Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

The object of this paper is so called adverbal prepositions in Russian; such as VOKRUG (kostra) ‘around smth.’, DALEKO OT (doma) ‘far from smth.’, etc. By definition, an adverbial preposition coincides with an adverb (cf. VOKRUG) or contains an adverb and a preposition (cf. DALEKO OT). In most cases, an adverbial preposition and the underlying adverb have the same meaning and the same semantic actants. The only difference between an adverbial preposition and the underlying adverb is the mode of expression of the main semantic actant. Cf. GOREL KOSTER, VOKRUG (preposition) KOSTRA STOJALI LIUDI ‘A fire was burning, people were standing around it’ vs GOREL KOSTER, VOKRUG (adverb) STOJALI LIUDI ‘A fire was burning, people were standing around’. Both the adverbial preposition VOKRUG and the adverb VOKRUG have a semantic actant ‘reference point’ and in both examples the word ‘fire’ expresses this actant. But the adverbial preposition governs this noun predicting its case-form and its linear position in a sentence. The adverb does not govern the noun; the only requirement is that this object must be already mentioned (so the noun must be somewhere in the preposition to the adverb). In this regard the adverbs under discussion are similar to connectors. Adverbial prepositions are easily described in the frameworks of valency theory. I argue that some refinements of valency theory are necessary for representing syntactic properties of underlying adverbs. I also demonstrate that it is more convenient to represent so called adverbal prepositions as adverbs but not as prepositions.

REGULARIZATION OF PROBABILISTIC TOPIC MODELS TO IMPROVE INTERPRETABILITY AND DETERMINE THE NUMBER OF TOPICS

Vorontsov K. V. (voron@forecsys.ru), Dorodnicyn Computing Centre of RAS; Moscow Institute of Physics and Technology, Moscow, Russia, **Potapenko A. A.** (anya_potapenko@mail.ru), Lomonosov Moscow State University, Moscow, Russia

Probabilistic topic modeling is a rapidly developing branch of statistical text analysis. The topic model uncovers a hidden thematic structure of the text collection. Learning a topic model from a document collection has an infinite set of solutions. The nonuniqueness results in a weak interpretability and instability of the solution. To tackle these problems we use a new multi-objective approach — Additive

Regularization of Topic Models (ARTM). ARTM is a non-Bayesian framework free of redundant probabilistic assumptions, which dramatically simplifies the inference of topic models and makes topic models easy to design, infer, and explain. With ARTM we combine four regularizers to concentrate common vocabulary words in background topics, to make domain topics sparse and distinct, and to eliminate insignificant topics. In our experiments the combination of the regularizers improves sparsity, coherence, purity, and contrast criteria at once almost without any loss of the perplexity.

WHY STANDARD ORTHOGRAPHY? BUILDING THE USTYA RIVER BASIN CORPUS, AN ONLINE CORPUS OF A RUSSIAN DIALECT

Waldenfels R. von (ruprecht.waldenfels@gmail.com), Instytut Podstaw Informatyki Polskiej Akademii Nauk, Warsaw, Poland, **Daniel M.** (misha.daniel@gmail.com), National Research University Higher School of Economics, Moscow, Russia, **Dobrushina N.** (nina.dobrushina@gmail.com), National Research University Higher School of Economics, Moscow, Russia

The paper describes a corpus of dialectal Russian speech under development. The corpus relies on interviews conducted by a joint Swiss-Russian team in the summer of 2013 in a small cluster of North Russian villages with the goal of studying the local dialect from a sociolinguistic and dialectological perspective.

The interviews are transcribed into standard Russian and thus do not involve a detailed phonetic representation. The text is then lemmatized and grammatically annotated with standard tools and fed into a corpus. The corpus can be queried via a web-based interface which provides the user with access to the original sound recordings on a per-utterance level. This design, the paper argues, allows for a rapid development of the corpus without a major loss in usability, since the audio data are readily available. Future plans include more field trips as well as a more convenient interface providing, among other features, for user correction of the transcription.

CORPUS AND INSTRUMENTAL METHODS IN ANALYSING FICTION AUDIO RECORDINGS

Yanko T. E. (tanya_yanko@list.ru), Institute of Lingustics of the Russian Academy of Sciences, Moscow, Russia

This paper aims at analyzing the communicative structure of sentences with new information placed sentence-initially. The point of departure is the analysis of the examples from Pushkin and L. Tolstoy discussed in [Kovtunova 1979]. Russian linguists traditionally used Russian classical texts for verifying and exemplifying their scientific hypotheses. Currently, a vast amount of fiction read by the best actors and the development of convenient systems of spoken speech analysis, such as Praat or Speech Analyzer, open an easy access to the prosodic structure of a sentence. The prosodic structure, in its turn, allows of modelling the communicative structure, since prosody is the main means to manifest the communicative division of a sentence into theme and rheme. Availability of modern corpora and tools for investigation demonstrate the prosodic and the corresponding communicative structures really employed by the speakers who voiced over the texts of Pushkin and L. Tolstoy, and who presumably never read I. I. Kovtunova's papers. For investigation, a minor working corpus of sounding texts was set up. The new tools completely confirmed the basic I. I. Kovtunova's findings obtained in the second half of the 20th century by the method of introspection. Nevertheless, some new additions acquired by the use of the new sources of material and the corpus techniques correct the results achieved in [Kovtunova 1979] and substantially widen the variety of theme-rheme structures applicable to the sentences with new information placed sentence-initially.

RUSSIAN VERBAL ASPECT AND MACHINE TRANSLATION

Zangenfeind R. (r.zangenfeind@lmu.de), **Sonnenhauser B.** (basonne@lmu.de), University of Munich, Munich, Germany

Rule-based machine translation still offers some very beneficial facets for linguistic theory, because by implementing rules on the computer linguistic theory can be verified in practice. One of the most intricate problems for machine translation is grammatical aspect in Russian when

it has to be translated into a language either lacking aspect or having a different aspect system. On the categorical level, aspect has only approximate equivalents in non-Slavic languages, such as the progressive form in English, for instance. In addition, language-internally, its semantics and interpretation cannot be sufficiently captured with only one specific characteristic feature. In this paper, we aim at establishing a basis for the machine translation of the Russian aspect. To do so, we discuss an approach to describe the interaction of verb and aspect semantics in a systematic way. Moreover, we describe a possible annotation for the aspectual information that is provided by further lexical components contributing to the meaning computation. This allows for the formulation of rules for machine translation into target languages where the grammatical category of aspect is realized differently or not present at all.

SENTENTIONAL ARGUMENTS AND EVENT STRUCTURE

Zimmerling A. V. (fagraey64@hotmail.com), Sholokhov Moscow State University for the Humanities, Moscow, Russia; Institute of Linguistics of the Russian Academy of Sciences, Moscow, Russia

This paper is addressed the interaction of subject marking and event structure in languages, which allow sententional arguments in the subject position. In Russian and other Slavic languages sententional subjects share a number of formal and semantic properties with zero subjects with role-and-reference features and with so called oblique subjects, i.e. subject-like arguments marked with an oblique case. I argue that sententional subjects represented by bare that-clauses (Rus. *čto*-clauses) cannot have the roles of Agent/Causer, while zero subjects can. I also argue that the capacity of taking *to, čto* P- clauses, i.e. that-clauses headed by a correlative pronoun *to* serves as diagnostics for a number of verbal classes. Causative predicates like *vynudit', zastavit', sklonitj k čemu-l.* only take *to, čto P-* clauses, but not bare *čto P-* clauses as surface subjects. Factive predicates like *znat', razdražat'* etc. take *to, čto P-* clauses, but not bare *čto P-*clauses as surface subjects while non-factive predicates like *dumat', mereščit'sa* only take bare *čto P-* clauses. Nominal predicatives forming Dative-Predicative-Structures (DPS) with an oblique subject marked with dative case and specified as {+ animate; + referential} split into two groups. Russian DPS predicatives from the *stydno, dosadno, protivno, vse ravno* group only take bare *čto P-* clauses and invariably behave as non-factive verbs in all contexts with an overt oblique subject. Russian DPS predicatives from the *izvestno, neizvestno, stranno, bezrazlično* group both take bare *čto P*-clauses and headed *to, čto P*-clauses, i.e. can be used in factive contexts as well. That means that their sententional argument can both get the status of a fact i.e. verified proposition P, logical truth, and an intentional situation, e.g. subjective evaluation of P, inner vision of P etc. Russian has two expletive elements—*eto* and *to,* but their syntax is different. *Eto* behaves as surface subject of the matrix clause and alternates with oblique subjects and sententional arguments in the subject position while *to* cannot be separated from the complement clause and reaches the subject position only in combination with the CP.

ILYA SEGALOVICH AND DEVELOPMENT OF IDEAS OF COMPUTATIONAL LINGUISTICS TO YANDEX

Zelenkov Yu. G. (yuryz@yandex-team.ru), **Zobnin A. I.** (alzobnin@yandex-team.ru), **Maslov M. Yu.** (maslov@yandex-team.ru), **Titov V. A.** (uht@yandex-team.ru), Yandex, Moscow, Russia

In the article the most important and interesting linguistic projects led by Ilya Segalovich (1964–2013) — one of the founders of the Yandex search engine — are considered. He also took part in their development. The following projects are among them. Development of the morphological analysis and synthesis of Russian words with a possibility of processing «new» words not included in the dictionary; solving the problem of morphological ambiguity for the Russian language with the help of normalizing substitutions; practical transcription of foreign, individual and common words; automatic positioning of stresses and the analysis of poetic texts; creation of efficient methods of recognizing fuzzy duplicates for textual documents; development of the information and require system «The National Corpus of Russian», etc. Key ideas and approaches connected with the searching of solutions to complicated linguistic problems are described, and Ilya's role in the invention of these approaches and their further development is stated. Examples of nontrivial linguistic algorithms developed by Ilya in collaboration with his colleagues are given.

Авторский указатель

Author Index

Научное издание

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной
Международной конференции «Диалог»

Выпуск 13 (20). 2014

Ответственный за выпуск **А. А. Белкина**
Вёрстка **К. А. Климентовский**

Подписано в печать 13.05.2014
Формат 152 × 235
Бумага офсетная
Тираж 250 экз. Заказ № 75

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