Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной международной конференции «Диалог» (2018)

Выпуск 17

Computational Linguistics and Intellectual Technologies

Papers from the Annual International Conference “Dialogue” (2018)

Issue 17

УДК 80/81; 004
ББК 81.1
К63

Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 30 мая — 2 июня 2018 г.). Вып. 17 (24), 2018.

Сборник включает 64 доклада международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2018», 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

17-й выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит избранные материалы 24-й международной конференции «Диалог». На основании мнений наших рецензентов для публикации в ежегоднике Редсоветом были отобраны 64 доклада из числа примерно ста работ, которые были рекомендованы по результатам рецензирования для представления на конференции в 2018 году.

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

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

Все направления Диалога важны, но каждый год какие-то темы занимают особое место в программе конференции и в составе ежегодника. В этом году главными темами «Диалога» станут:

- Лингвистическая интерпретация результатов глубокого машинного обучения («черных ящиков»). Принципиальная особенность невероятно популярных сегодня многослойных нейронных сетей заключается в том, что мало кто понимает, какие факторы на самом деле повлияли на итог их работы. Они являются своеобразными «черными ящиками». Это, с одной стороны, порождает некоторый пессимизм среди части лингвистов относительно перспектив их науки, а с другой — заставляет задуматься о возможности использования глубокого обучения для собственно лингвистических исследований, о содержательной интерпретации его результатов. Возможно, такое понимание поможет улучшить и работу самих нейронных сетей.
- Методы применения технологий анализа больших данных к задачам, для которых таких данных не хватает. Не секрет, что для многих задач текстовой аналитики очень сложно найти адекватные обучающие датасеты. А решать их хочется. Существуют разные подходы к этой проблеме,

от автоматической генерации обучающих данных до технологий переноса результатов глубокого обучения с задач, где данных много, на задачи, где их не хватает (т. н. TransferLearning).
- Мультимодальная коммуникация. Это изучение всех сфер речевого акта, языка, интонации, мимики и жестов, эмоционального и коммуникативного поведения. На конференции обсудят новые результаты и возможности их практического применения, например, для диалоговых агентов или роботов.

Статьи в сборнике публикуются на русском и английском языках. При выборе языка публикации действует следующее правило:
- доклады по компьютерной лингвистике должны подаваться на английском языке. Это расширяет их аудиторию и позволяет привлекать к рецензированию международных экспертов.
- доклады, посвященные лингвистическому анализу русского языка, предполагающие знание этого языка у читателя, подаются на русском языке (с обязательной аннотацией на английском).

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

Мы обращаем внимание авторов и читателей сборника, что с этого года мы отказались от печати сборника на бумаге, поскольку бумажный вариант пользуется все меньшей популярностью. Сборник, как и в прошлые годы, размещается на сайте конференции и индексируется Scopus.

Программный комитет конференции «Диалог»

Редколлегия сборника «Компьютерная лингвистика и интеллектуальные технологии»

Организаторы

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

Учредителями конференции являются:

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

Международный программный комитет

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

Организационный комитет

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

Секретариат

Павлинова Ольга Сергеевна, *координатор оргкомитета*	Компания ABBYY
Ульянова Анна Вячеславовна, *секретарь оргкомитета*	РГГУ
Гусева Анна Александровна, *координатор Dialogue Evaluation*	Компания ABBYY
Севергина Екатерина Александровна, *администратор оргкомитета*	Компания ABBYY

Рецензенты

Антонова Александра Александровна
Азарова Ирина Владимировна
Андрианов Андрей Иванович
Апресян Валентина Юрьевна
Архангельский Тимофей Александрович
Баранов Анатолий Николаевич
Бенко Владимир
Богданов Алексей Владимирович
Богданова-Бегларян Наталья Викторовна
Браславский Павел Исаакович
Васильев Виталий Геннадьевич
Гершман Анатолий
Гращенков Павел Валерьевич
Губин Максим Вадимович
Даниэль Михаил Александрович
Добров Борис Викторович
Добровольский Дмитрий Олегович
Добрушина Нина Роландовна
Добрынин Владимир Юрьевич
Зализняк Анна Андреевна
Захаров Леонид Михайлович
Иванов Владимир Владимирович
Ильвовский Дмитрий Алексеевич
Иомдин Борис Леонидович
Иомдин Леонид Лейбович
Исаев Игорь Игоревич
Катинская Анисья Юрьевна
Клышинский Эдуард Станиславович
Кибрик Андрей Александрович
Князев Сергей Владимирович
Кобозева Ирина Михайловна
Копотев Михаил Вячеславович
Коротаев Николай Алексеевич
Котельников Евгений Вячеславович
Котов Артемий Александрович
Крейдлин Григорий Ефимович
Левонтина Ирина Борисовна
Леонтьев Алексей Петрович
Лобанов Борис Мефодьевич
Лукашевич Наталья Валентиновна
Лютикова Екатерина Анатольевна
Марков Александр Юрьевич
Мисюрев Алексей Владимирович
Наков Преслав
Недолужко Анна Юрьевна
Пазельская Анна Германовна
Паперно Денис Аронович
Панченко Александр Иванович
Переверзева Светлана Игоревна
Пиперски Александр Чедович
Подлесская Вера Исааковна
Смирнов Иван Валентинович
Селегей Владимир Павлович
Слюсарь Наталия Анатольевна
Сорокин Алексей Андреевич
Тихомиров Илья Александрович
Толдова Светлана Юрьевна
Урысон Елена Владимировна
Усталов Дмитрий Алексеевич
Федорова Ольга Викторовна
Хохлова Мария Владимировна
Циммерлинг Антон Владимирович
Шаврина Татьяна Олеговна
Шаров Сергей Александрович
Шелманов Артём Олегович

Contents*

* The reports of each section are ordered by the surname of the first author in compliance with the English alphabet.

INTRA-TEXT COHERENCE AS A MEASURE OF TOPIC MODELS' INTERPRETABILITY

Alekseev V. A. (wasya.alekseev@gmail.com),
Bulatov V. G. (bt.uytya@gmail.com),
Vorontsov K. V. (vokov@forecsys.ru)
Moscow Institute of Physics and Technology (State University)

The article is devoted to the problem of how to automatically measure the interpretability of topic models. Some new, intra-text, approaches to estimate the interpretability of the topics are proposed. Computational experiments are conducted with the use of text files from "PostNauka", which is a collection of popular science content.

Keywords: topic modeling, topic coherence, topic interpretability, text segmentation, topic model, PLSA, LDA, BigARTM, text analysis, machine learning

ВНУТРИТЕКСТОВАЯ КОГЕРЕНТНОСТЬ КАК МЕРА ИНТЕРПРЕТИРУЕМОСТИ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ ТЕКСТОВЫХ КОЛЛЕКЦИЙ

Алексеев В. А. (wasya.alekseev@gmail.com),
Булатов В. Г. (bt.uytya@gmail.com),
Воронцов К. В. (vokov@forecsys.ru)
Московский Физико-Технический Институт

Статья посвящена задаче измерения интерпретируемости и когерентности тематических моделей. Предлагается новый, внутритекстовый, подход к оценке меры согласованности темы. Вычислительные эксперименты проводятся на коллекции научно-популярного контента «ПостНаука».

Ключевые слова: тематическое моделирование, интерпретируемость, когерентность, сегментация, тематическая модель, PLSA, LDA, BigARTM, анализ текстов, машинное обучение

1. Introduction

Topic modeling is a text analysis method which aims to discover hidden thematic structure in large collections of texts. Topic models are used in information retrieval [10], documents' categorization [12], social networks' data analysis, [16, 15], recommendation systems [10, 7], exploratory search [5] and other areas. After the processing of documents' collection, a topic model gives a set of topics covered in the documents, the distribution of these topics in the documents, and words that characterize each topic [11].

The interpretability is a desirable property of a good topic model [19]. A topic is said to be well interpreted, if it corresponds to real-world concept of interest. However, the topics derived by topic models may not be clear and understandable, they may include words from different weakly related areas. [8]

Recently, an automated procedure estimating the interpretability was introduced. This method evaluates the list of the most frequent topic words and favorably compares to the human experts' judgements of the same list.

However, we believe that this approach suffers from several fundamental limitations. We argue that these limitations bring into question the common practice of treating coherence and interpretability as equivalent.

The aim of this paper is twofold. The first is to outline a class of issues inherent in a traditional notions of coherence. A key problem with this approach is that reducing the topic model to a short list of words loses too much precision. Previous studies linking coherence and interpretability failed to take this into account.

However, the proportion of text covered by these top frequent words is not controlled in any way. We show that in practice this proportion is too small to justify treating coherence and interpretability as equivalent.

The second purpose is to demonstrate the feasibility of the alternative approach which we call the *intra-text coherence*, defined as an average thematic similarity of terms, closely located in the text. To justify this new measure, we will adapt the procedure used in [3], [8] and [14].

2. Related work

For the topic modeling purposes, the *topic* is defined as a probability distribution over words. For example, the topic named "theatre" could be a probability distribution concen-trated on a words such as "actor", "play", "premiere", "parterre" and "spectator" (on the contrary, the probability of words such as "loan" and "vertebrates" would be extremely low or even zero).

The topic model can be described by two distributions: $\phi_{wt}=(w|t)$, the probability to draw a word w from the topic t and $\theta_{td}=p(t|d)$, the probability to find a topic t within the document d.

Early work on topic modeling conceptualized it as an intermediate stage of information retrieval pipeline. The possibility of meaningful interpretation was an afterthought. For measuring the quality of topics when evaluated against human judgments, several metrics were proposed.

Currently, there is a consensus among researchers that the evaluation of human interpretability should conform to the following framework:

1) Picking some small set of words for each topic (typically, a list of ten most frequent words, but the more sophisticated approaches are possible [2]). The term *top tokens* has come to be used to refer to this set.

2) a. Presenting this set to a human expert to obtain a human judgment of a set quality.

or

b. Gathering an array of co-occurrence statistics associated with members of this set and performing a series of calculations involving these numbers.

This framework was introduced in seminal works of Blei [14, 3] and Mimno [8] and then greatly developed by the topic modeling community. We will call this extensive category of metrics *top-tokens based*.

The main attraction of top-token based measures is their simplicity. Instead of evaluating the whole probability distribution, the researcher only has to look at the short list of the most "representative" words.

However, their inherent limitation is deeply rooted in the same thing. The list of top five to ten words reflects only part of the whole probability distribution, and poorly (if at all) characterizes how good topic model does represent the particular corpus.

We argue that the list of the most frequent words is inadequate in justifying the quality of topic model regardless of the method of its analysis. This applies equally to the human experts' ratings and the automated procedures based on the word co-occurrence counts.

3. Towards a better interpretability metric

As was previously noted, traditional coherence metrics consist of two steps: first, they use information from $(w|t)$ distribution; secondly, they retrieve the co-occurrence statistics.

The idea behind automated coherence measures is to find out how often do certain words appear together within the sliding context window and compare that number to the frequency predicted by pure coincidence. The topic is said to be coherent if the positions of its words tend to cluster, do not appear to be random.

This is reminiscent of the linguistic phenomenon of textual cohesion [1]: the sentences of natural language texts are connected to each other via syntactic and lexical devices such as word repetition, synonyms/near-synonyms, hyponyms and so on.

We conjecture that the natural language texts are divided into coherent spans which contain only small number of latent topics. According to this assumption, the purpose of topic modeling should be understood as an adequate segmentation of the initial text into thematically homogeneous fragments consisting of a handful of topics.

Note that frequent top-words co-occurrences is an indirect sign that the topic is represented in the text collection as a coherent text fragment.

Therefore, we argue that interpretability of a topic should be evaluated not only by the consistency of top-words use, but also by the consistency of all topic words use within text segments. We could obtain an automated measure of the model interpretability by examining the degree the topic model violates this consistency.

Instead of drawing inferences about the whole topic based on behaviour of the short list of ten most frequent words, one should start by examining words appearing together in a text and then proceed by comparing their $(t|w,d)$.

This procedure will be dealt with in more detail in the following section.

4. Coherences

In this paper, we present several automatic measures distinct from traditional *top-token based* approaches.

The first method—SemantiC (Semantic Closeness)—estimates semantic proximity of closely located in the text words as vectors with components $(t|w)$. To estimate the proximity between words one can calculate l2 distance between the corresponding vectors

$$SemantiC_{L2}|_t = -\langle [\rho(\boldsymbol{w}_i, \boldsymbol{w}_j) \leq window] \|\boldsymbol{w}_i - \boldsymbol{w}_j\|_2 \rangle$$

where $\rho(\boldsymbol{w}_i, \boldsymbol{w}_j)$—text-distance between words (number of other words between them), window—window of words, in which $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ are considered to be close in text-distance. Minus sign makes coherence higher if words' vectors are close. In addition to the Euclidean distance, Cosine Similarity measure can be used:

$$SemantiC_{Cos}|_t = +\langle [\rho(\boldsymbol{w}_i, \boldsymbol{w}_j) \leq window] cos(\boldsymbol{w}_i, \boldsymbol{w}_j) \rangle$$

The third proposed way to estimate semantic closeness by topic is to calculate variance between components corresponding to this topic:

$$SemantiC_{Var}|_t = Varianc(\boldsymbol{w}_i(t), \boldsymbol{w}_{i+1}(t), \ldots, \boldsymbol{w}_{i+window}+(t))$$

Before computing, all vectors were multiplied by 1000, so as to increase the result value for the coherence.

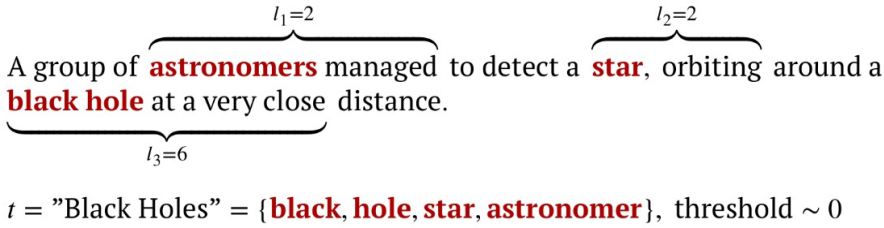


Figure 1. An example illustrating the idea of TopLen coherence

As long as words of a topic under interest are observed, they are counted. If some unrelated word is encountered it is also counted but gives a negative penalty. When the absolute value of total penalty appears to be quite big, the process stops, and the number of counted words gives one value of topic length.

Another method—TopLen (Topic Length)—calculates the average duration of the topic in text. The auxiliary function score (w_j,t) returns the difference between the component of the vector corresponding to the topic and the maximal component among the other topics. Non-negative parameter threshold smooths the effect when TopLen encounter words not from the topic while counting topic length, the process of counting continues as long as threshold (chosen to be 0.01) plus sum of scores is non-negative (see Figure 1 for an example).

The last proposed method—FoCon (Focus Consistency)—evaluates how much differ adjacent words throughout the whole text, summing the pairs of differences between corresponding components of $(t|w)$ vectors (components, by means of which the differences are calculated, are the maximal components of the adjacent words vectors). Minus sign serves the same role as in case of SemantiC—coherence rises when words differ less.

$$FoCon|_t = -\sum_{d\in D}\sum_{\substack{w_i,w_j\in W_d\\ j-i=1}} \left|w_i[t_1]-w_j[t_1]\right| + \left|w_i[t_2]-w_j[t_2]\right|$$

5. Experiments

5.1. Interpretation and representation

Automated coherence measures rest on the word co-occurrence counts. If top tokens often appear together within the context window, this set of words is said to be *coherent*, i.e. these words fit together in a natural or reasonable way.

It is implicitly assumed that if set of top tokens is coherent, then the whole topic is coherent as well. Such arguments were criticized before [21], but we wish to understand the issue quantitatively. What fraction of collection is represented in the co-occurrence counts related to the given top token set?

Let Q be a set of words. We will call the position of word $w \in Q$ *represented* if it has a non-zero contribution to the Q co-occurrence counts (see Figure 2). We will measure the *representational frequency* of two topic models.

Our primary dataset is a corpus consisting of articles published in "PostNauka", a popular Russian online magazine about science. We investigate a topic model consisting of 19 subject-related topics and a single background topic (see Figure 3).

Напротив, если предположить существование суперсимметрии, то введение новых **частиц** приводит как раз к такому объединению. Оказывается, что суперсимметрия не только обеспечивает объединение взаимодействий, но и стабилизирует объединённую теорию, в которой присутствуют два совершенно разных масштаба: масштаб масс обычных **частиц** (порядка 100 масс протона) и масштаб великого объединения (порядка 10^{16} масс протона). Последний масштаб уже близок к так называемому планковскому масштабу, равному обратной ньютоновской константе тяготения, что составляет порядка 10^{19} масс протона. На этом масштабе мы ожидаем проявление эффектов квантовой гравитации. В этом моменте нас ожидает приятный сюрприз. Дело в том, что гравитация всегда стояла несколько особняком по отношению к остальным взаимодействиям. Переносчик гравитации, гравитон, имеет спин 2, в то время как переносчики остальных взаимодействий имеют спин 1. Однако суперсимметрия перемешивает спины.

first top words of *topic 3: физика* with **top 10** in bold: **частица**, **электрон**, **кварк**, **атом**, **энергия**, **вселенная**, **фотон**, **физика**, **физик**, **эксперимент**, масса, теория, свет, симметрия, протон, эйнштейн, нейтрино, вещество, квантовый, ускоритель, детектор, волна, эффект, свойство, спин, гравитация, материя, адрон, поль, частота

Figure 2: Words used to calculate coherence. We see a single top token ("частиц") and a wide range of weakly topical words, which are ignored while calculating coherence by the traditional methods.

Topic	First Top-Word	Second Top-Word	Third Top-Word
1: математика	математика (0.016)	задача (0.008)	декарт (0.008)
2: технологии	технология (0.015)	робот (0.012)	сеть (0.010)
3: физика	частица (0.027)	электрон (0.015)	кварк (0.015)
4: химия	химия (0.021)	молекула (0.019)	материал (0.016)
5: земля	земля (0.029)	планета (0.028)	атмосфера (0.012)
6: астрономия	звезда (0.039)	галактика (0.031)	вселенная (0.019)
7: биология	клетка (0.027)	организм (0.011)	мозг (0.010)
8: медицина	пациент (0.016)	препарат (0.012)	заболевание (0.012)
9: психология	психология (0.009)	мозг (0.009)	психолог (0.008)
10: экономика	экономика (0.016)	страна (0.010)	цена (0.008)
11: история	история (0.010)	историк (0.007)	власть (0.006)
12: политика	государство (0.014)	политика (0.012)	политический (0.011)
13: социология	социология (0.013)	социолог (0.009)	социальный (0.008)
14: культура	культура (0.015)	фильм (0.007)	искусство (0.006)
15: образование	университет (0.021)	образование (0.014)	школа (0.013)
16: язык	язык (0.077)	слово (0.037)	словарь (0.011)
17: философия	философия (0.018)	философ (0.013)	философский (0.008)
18: религия	святилище (0.010)	религия (0.007)	царь (0.006)
19: россия	россия (0.028)	страна (0.009)	русский (0.009)

Figure 3: PostNauka's topics, each represented by its 3 top-words

Next, we will focus on the topic model presented in [3], which uses a sample of Wikipedia articles. This model was identified as best based on assessment of top 10 tokens by human experts. This model consists of 50 topics.

As can be seen from the table 1, top-tokens cover a vanishing fraction of corpus. Informally speaking, top token-based measures ignore more than 98% of the collection!

Table 1: The proportion of corpus contributing to the co-occurrence counts of top 10 most frequent words for each topic

	PostNauka	Wikipedia
Minimum	0.000159	0.000065
Median	0.000483	0.000293
Mean	0.000619	0.000356
Maximum	0.002764	0.001149
Total	0.012027	0.016585

5.2. Ground truth

The evaluation of interpretability is extremely labor-intensive. The strength of top token-based measures is their ability to reduce topics of the topic model to the accessible list of words. Even then, gathering human judgments about a large number of topics is a daunting task.

This leaves us with a difficult problem. On one hand, we try to construct a measure taking into account the whole Φ and Θ matrices and the whole corpus. On the other hand, validating such measure requires comparing them to the human judgment. Therefore, one needs to somehow obtain human ratings about the whole corpus and the whole probability distribution.

We propose a way to circumvent this infeasible procedure: instead of asking human experts to produce a number of labels, we generate a semi-synthetic dataset with known labels. In this enterprise, the structure of PostNauka dataset is of a tremendous help. The topics of articles are general and diverse enough to make the majority of documents *monotopical*: i.e. every word of such document could be attributed either to a single specific topic or to background topic.

We use these monotopical documents to produce a semi-synthetic dataset. The idea is to "cut" the monotopical documents into smaller monotopical segments and then "sew" them together in random order. The intent of this semi-synthetic dataset is to serve as a ground truth by which topic models can be evaluated

The generation procedure ensures that we know true topic labels for every word. Given this information, it is possible to define segm to be the segmentation quality of any topic model. There are two natural ways to do this:

- soft: for each topic t the sum of $p(t \mid d, w)$ on all pairs (d, w), $d \in D, w \in W_d$ is calculated, with total result equals to the sum of these sums for all topics
- strict: for each topic t for all segments of topic t the number of coincidences of topic, predicted by the model for a word in a document, with the topic t of the segment to which this word belongs $[\operatorname{argmax}_\tau p(\tau \mid d, w) = t]$.

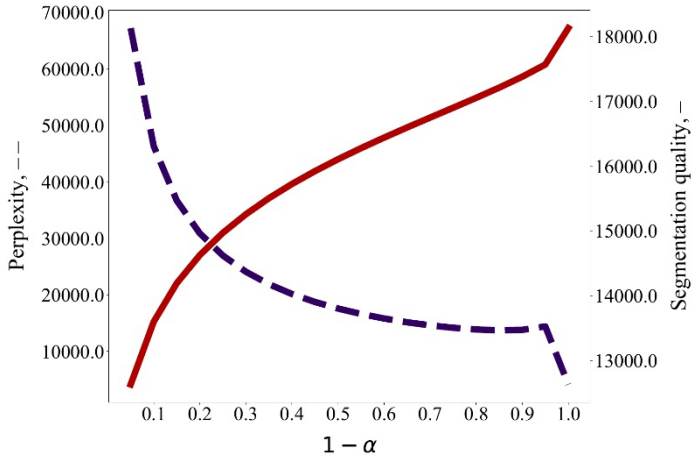


Figure 4. The relationship between segmentation quality and perplexity of topic model

On the X axis is the proportion of good Phi matrix: one minus α (degree of Phi degradation). The fact that segmentation quality monotonically increases when perplexity decreases implies that the proposed segmentation quality may be used as a measure of quality of topic models.

Having established the ground truth, we are able to evaluate different coherence measures. The quality of each candidate measure coh is defined to be a Spearman correlation coefficient between the function value and the segmentation quality.

For this purpose, we generated a number of different Φ matrices as a weighted combination of Φ_{good} (the topic model of PostNauka dataset, discussed above) and Φ_{bad} (a set of random columns taken from Dirichlet $(0.01^{|W|})$ distribution):

$$(\alpha)=\alpha\cdot\Phi_{bad}+(1-\alpha)\Phi_{good}$$

For each α, the segmentation quality and all the investigated coherence metrics were calculated. Thus, a sample $\{\langle \text{soft}(m), \text{strict}(m), c_1(m), c_2(m), \ldots c_n(m)\rangle \mid m \in M, c_i \in \text{Coh}, 1 \leq i \leq |\text{Coh}|\}$ was obtained. Four series of experiments were conducted, with different Φ_{bad} matrices.

Good Topic Model

topic 16: язык
Категория будущего времени в большинстве языков Африки отсутствует. Есть много способов говорить о будущем, но это более сложные способы, касающиеся предположения, желания. Нормальный африканский грамматический приём — не говорить "я это сделаю"или "это будет а сказать "это возможно"или "я хочу это сделать"они говорят о будущем, но "попадают"в будущее непрямым путём.

topic 12: политика
И я посылаю деньги борцам за независимость Курдистана, участвую в акциях поддержки курдских повстанцев и так далее. Вот такое наложение друг на друга разных членств, разных "гражданств". В литературе последних десяти лет бытуют такие выражения, как "гендерное гражданство "экономическое гражданство". Первое указывает на членство в воображаемом сообществе женщин, приверженных идеям феминизма.

SQ (S)	SQ (H)	N	M	SC L2	SC Cos	SC Var	TL	FC
16.0e3	**3.76e4**	**-3.65**	**-2.69**	**-3.70**	0.700	**-8.12e3**	**3.45**	**-5.44e4**

Bad Topic Model

topic 16: язык
Категория будущего времени в большинстве языков Африки отсутствует. Есть много способов говорить о будущем, но это более сложные способы, касающиеся предположения, желания. Нормальный африканский грамматический приём — не говорить "я это сделаю"или "это будет а сказать "это возможно"или "я хочу это сделать"они говорят о будущем, но "попадают"в будущее непрямым путём.

topic 12: политика
И я посылаю деньги борцам за независимость Курдистана, участвую в акциях поддержки курдских повстанцев и так далее. Вот такое наложение друг на друга разных членств, разных "гражданств". В литературе последних десяти лет бытуют такие выражения, как "гендерное гражданство "экономическое гражданство". Первое указывает на членство в воображаемом сообществе женщин, приверженных идеям феминизма.

SQ (S)	SQ (H)	N	M	SC L2	SC Cos	SC Var	TL	FC
5.54e3	1.10e4	-4.83	-3.12	-12.9	0.947	-37.0e3	2.87	-13.9e4

Figure 5. Illustration of a model segmentating semisynthetic text

The figure shows two segments of size 50 words from different topics after being processed by *Bad Topic Model* or *Good Topic Model* (discussed above). These segments were extracted from one of the generated documents, in which they were adjacent. Words that are not labeled were assigned topics different from the two of represented segments. Below the segments are coherence values. SQ (S)—stands for soft segmentation quality, SQ (H)—strict segmentation quality, N—Newman, M—Mimno, SC—SemantiC, TL—TopLen, FC—FoCon. Values in bold indicate that coherence function rises as model's quality increases.

Table 2: Spearman correlations between coherences and segmentation qualities (soft) for datasets with different sizes of segments: 50, 100, 200 and 400 words and with 5 topics in each document

Coh	Corr	Coh	Corr	Coh	Corr	Coh	Corr
Newman	0.75	Newman	0.94	Newman	0.80	Newman	0.85
Mimno	0.96	Mimno	0.96	Mimno	0.94	Mimno	0.97
SC L2	0.92	SC L2	0.91	SC L2	0.70	SC L2	0.59
SC Cos	−0.97	SC Cos	−0.96	SC Cos	−0.97	SC Cos	−0.96
SC Var	**1.00**	SC Var	**1.00**	SC Var	**1.00**	SC Var	**1.00**
TopLen	**1.00**	TopLen	**1.00**	TopLen	**1.00**	TopLen	**1.00**
FoCon	**1.00**	FoCon	**1.00**	FoCon	**1.00**	FoCon	**1.00**

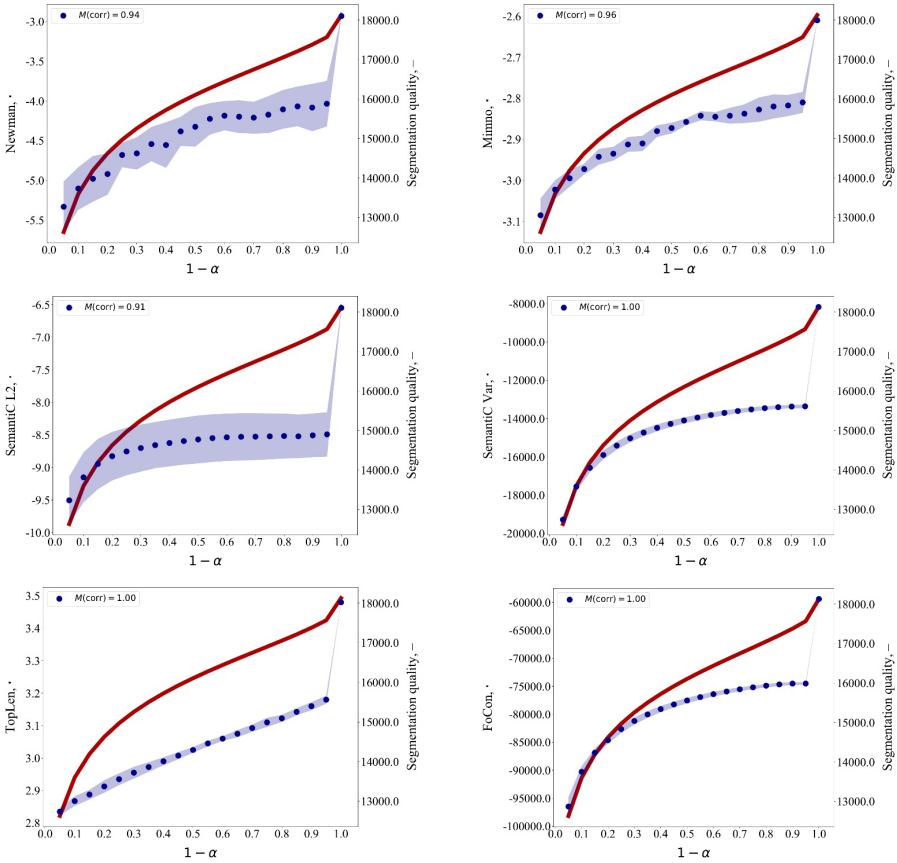


Figure 6. The comparison of different coherence measures with segmentation quality as a function of α, the topic model degradation parameter. Coherence values drawn on the plots are averaged values from 4 series (α) which differ in Φ_{bad} matrix

6. Results

Three new methods for estimating topic model's interpretability are presented: SemantiC, TopLen and FoCon,—which try to take into account all words of the text when evaluating coherence. The new methods show that this is possible to develop an indicator of interpretability able to overcome the shortfalls of top token-based measures.

Experiments on semisynthetic dataset, consisting of segments of different topics, were conducted in order to analyze some properties of new coherences and existing ones.

Proposed methods demonstrate high correlations with the quality of semisynthetic dataset segmentation. SemantiCVar and TopLen appear to perform best.

Acknowledgments

The work was supported by Government of the Russian Federation (agreement 05.Y09.21.0018) and the Russian Foundation for Basic Research grant 17-07-01536. We thank Alexander Romanenko and Irina Efimova for their assistance in data collection.

All experiments with the data were carried out with the use of the BigARTM library [9, 18].

References

1. *Harold W. Kuhn.* "The Hungarian method for the assignment problem". In: Naval Research Logistics (NRL) 2.1–2 (1955), pp. 83–97.
2. *David M. Blei and John D. Lafferty.* "Topic models". In: Text mining: classification, clustering, and applications 10.71 (2009), p. 34.
3. *Jonathan Chang et al.* "Reading Tea Leaves: How Humans Interpret Topic Models". In: Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7–10 December 2009, Vancouver, British Columbia, Canada. Ed. by Yoshua Bengio et al. Curran Associates, Inc, 2009, pp. 288–296. isbn: 9781615679119.
4. *David Newman, Sarvnaz Karimi, and Lawrence Cavedon.* "External evaluation of topic models". In: Australasian Document Computing Symposium, December 2009. 2009, pp. 11–18.
5. *Ianina A., Golitsyn L., Vorontsov K.* Multi-objective topic modeling for exploratory search in tech news. In: Artificial Intelligence and Natural Language. AINL 2017, St. Petersburg, Russia, September 20–23, 2017. Ed. by Filchenkov A., Pivovarova L., Žižka J. Communications in Computer and Information Science, vol 789. Springer, Cham, 2017.—pp 181–193.
6. *Robert K Nelson.* "Mining the dispatch". In: Mining the dispatch (2010). url: http://dsl. richmond.edu/dispatch/pages/intro.
7. *Sang Su Lee, Tagyoung Chung, and Dennis McLeod.* "Dynamic Item Recommendation by Topic Modeling for Social Networks". In: Eighth International Conference on Information Technology: New Generations, ITNG 2011, Las Vegas, Nevada, USA, 11–13 April 2011. Ed. by Shahram Latifi. IEEE Computer Society, 2011, pp. 884–889. isbn: 978-0-7695-4367-3.

8. *David Mimno et al.* "Optimizing Semantic Coherence in Topic Models". In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP '11. Edinburgh, United Kingdom: Association for Computational Linguistics, 2011, pp. 262–272. isbn: 978-1-937284-11-4.
9. *Kochedykov D., Apishev M., Golitsyn L., Vorontsov K.* Fast and Modular Regularized Topic Modelling. In: Proceeding Of The 21St Conference Of FRUCT (Finnish-Russian University Cooperation in Telecommunications) Association. The seminar on Intelligence, Social Media and Web (ISMW). Helsinki, Finland, November 6–10, 2017. Pp.182–193.
10. *Chong Wang and David M. Blei.* "Collaborative topic modeling for recommending scientific articles". In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21–24, 2011. Ed. by Chid Apté, Joydeep Ghosh, and Padhraic Smyth. ACM, 2011, pp. 448–456. isbn: 978-1-4503-0813-7.
11. *David M. Blei.* "Probabilistic topic models". In: Commun. ACM 55.4 (2012), pp. 77–84.
12. *Timothy N. Rubin et al.* "Statistical topic models for multi-label document classification". In: Machine Learning 88.1–2 (2012), pp. 157–208.
13. *Nikolaos Aletras and Mark Stevenson.* "Evaluating topic coherence using distributional semantics". In: Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)—Long Papers. 2013, pp. 13–22.
14. *Jey Han Lau, David Newman, and Timothy Baldwin.* "Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality." In: EACL. 2014, pp. 530–539.
15. *Julio Cesar Louzada Pinto and Tijani Chahed.* "Modeling Multi-topic Information Diffusion in Social Networks Using Latent Dirichlet Allocation and Hawkes Processes". In: Tenth International Conference on Signal-Image Technology and Internet-Based Systems, SITIS 2014, Marrakech, Morocco, November 23–27, 2014. IEEE Computer Society, 2014, pp. 339–346. isbn: 978-1-4799-7978-3.
16. *Devesh Varshney, Sandeep Kumar, and Vineet Gupta.* "Modeling Information Diffusion in Social Networks Using Latent Topic Information". In: Intelligent Computing Theory—10th International Conference, ICIC 2014, Taiyuan, China, August 3–6, 2014. Proceedings. Ed. by De-Shuang Huang, Vitoantonio Bevilacqua, and Prashan Premaratne. Vol. 8588. Lecture Notes in Computer Science. Springer, 2014, pp. 137–148. isbn: 978-3-319-09332-1.
17. *Michael Röder, Andreas Both, and Alexander Hinneburg.* "Exploring the space of topic coherence measures". In: Proceedings of the eighth ACM international conference on Web search and data mining. ACM. 2015, pp. 399–408.
18. *Konstantin Vorontsov et al.* "BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections". In: Analysis of Images, Social Networks and Texts—4th International Conference, AIST 2015, Yekaterinburg, Russia, April 9–11, 2015, Revised Selected Papers. Ed. by Mikhail Yu. Khachay et al. Vol. 542. Communications in Computer and Information Science. Springer, 2015, pp. 370–381. isbn: 978-3-319-26122-5.

19. *Potapenko A. A., Popov A. S., Vorontsov K. V.* Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. In: Artificial Intelligence and Natural Language. AINL 2017, St. Petersburg, Russia, September 20–23, 2017. Ed. by Filchenkov A., Pivovarova L., Žižka J. Communications in Computer and Information Science, vol 789. Springer, Cham, 2017.— pp. 167–180.
20. *Halliday M. A. K., Hasan R.* Cohesion in English.—Routledge, 2014.
21. *Benjamin M. Schmidt.* "Words alone: Dismantling topic models in the humanities". In: Journal of Digital Humanities 2.1 (2012), pp. 49–65

IMPROVING PART-OF-SPEECH TAGGING VIA MULTI-TASK LEARNING AND CHARACTER-LEVEL WORD REPRESENTATIONS

Anastasyev D. G. (daniil_an@abbyy.com),
Gusev I. O. (ilya.gusev@phystech.edu),
Indenbom E. M. (eugene_i@abbyy.com)

ABBYY, Moscow Institute of Physics and Technology, Moscow, Russia

In this paper, we explore the ways to improve POS-tagging using various types of auxiliary losses and different word representations. As a baseline, we utilized a BiLSTM tagger, which is able to achieve state-of-the-art results on the sequence labelling tasks. We developed a new method for character-level word representation using feedforward neural network. Such representation gave us better results in terms of speed and performance of the model. We also applied a novel technique of pretraining such word representations with existing word vectors. Finally, we designed a new variant of auxiliary loss for sequence labelling tasks: an additional prediction of the neighbour labels. Such loss forces a model to learn the dependencies inside a sequence of labels and accelerates the process of training. We test these methods on English and Russian languages.

Keywords: pos-tagging, morphological analysis, deep learning, auxiliary loss, word representations

УЛУЧШЕНИЕ МОРФОЛОГИЧЕСКОГО ПАРСЕРА С ПОМОЩЬЮ ВСПОМОГАТЕЛЬНЫХ ЗАДАЧ ОБУЧЕНИЯ И ПРЕДСТАВЛЕНИЙ СЛОВ НА СИМВОЛЬНОМ УРОВНЕ

Анастасьев Д. Г. (daniil_an@abbyy.com),
Гусев И. О. (ilya.gusev@phystech.edu),
Инденбом Е. М. (eugene_i@abbyy.com)

ABBYY, Московский Физико-Технический Институт, Москва, Россия

1. Introduction

A machine learning model, which just learns by heart the train examples, is hardly adequate for most purposes. Good model should rather generalize across these examples. Regularization is one of the most useful tricks, which forces models to generalize better. However, simple regularization of the model's weights usually just reduces overfitting on the train data.

In the last few years, multi-task learning with auxiliary losses became increasingly popular as a method to improve generalization achieved by the model. In this scenario, additional objectives are used during the model training. Consequently, the model's parameters are shared between different tasks and the model learns more general representations from the train dataset.

Good features are even more important for the machine learning models. In NLP tasks meaningful word representations can become such features. In most cases, we have an access to the vast unlabelled data (mostly crawled from the Internet) and by several orders smaller amount of labelled data, specific to our task. The word2vec and similar frameworks give an opportunity to pretrain word vectors on the unlabelled data. Such pretrained vectors usually improve the performance of most NLP models.

Another way to obtain the words' vectors is to use their character-level representations. Such models are usually smaller than the models with word embeddings and they do not suffer from an inability to build vector for an out-of-vocabulary word.

In this work, we discuss the ways to improve the quality of part-of-speech tagging using the auxiliary losses and propose a new variant of character-level word representations.

2. Baseline Model

Part-of-speech (POS) tagging is the task of assigning each word in the given text an appropriate grammatical value. Various tasks in the field of natural language processing are using the results of POS tagging. Practically all modern POS-taggers are based on recurrent neural networks (RNN). The main reason is the ability of RNNs to handle long context dependencies. It means that in contrast to more classic models where the prediction is conditioned on the narrow context window the prediction made by the RNNs is based on the whole sentence.

In case of sequence labelling problems, even more useful is the usage of Bidirectional LSTM (BiLSTM): it outputs just a concatenation of forward and backward passes of ordinary LSTMs. BiLSTMs help to use both left and right contexts information, which is usually essential for the correct analysis.

Many works showed the superiority of BiLSTMs for sequence labelling tasks. Therefore, we considered the architecture of neural network with BiLSTM in the core as our baseline. The purpose of our work was to build the best possible set of features (words' representations) for the BiLSTM and to design better objective to improve the learning process.

3. Word Representations

3.1. Word-Level Representation

Frameworks like GloVe or word2vec can build an embedding matrix with meaningful word vectors in each row. Therefore, we can map a word to the corresponding one-hot-encoding vector in very high dimensional space and multiple the embedding matrix by this vector to obtain a low-dimensional dense representation of the word.

However, the embedding matrices are typically very big: for example, the 300-dimensional embeddings of 100 thousand words have 30 million parameters. Thus, the model's whole size becomes sometimes inappropriately large.

Another problem with the word-level representations is the fixed dictionary. We are able to deal only with the words from the embeddings' vocabulary while all other words are mapped into a single unknown word vector. Apparently, we cannot store vectors for every single word: besides some rare and novel words, there always can be found some misspelt words in most texts.

3.2. Character-Level Representation

In order to achieve an open vocabulary and have an ability to process misspellings, one can represent word not as a single number (or one-hot-encoding vector) but as a sequence of its letters. Then some character-level function should be applied to the sequence. This function has to map an arbitrary-length sequence to a fixed dimensional vector, which can be treated as the word's representation.

The most common way to deal with such sequences of letters is to use BiLSTM. In this case, we need only the last states from both forward and backward LSTMs (Fig. 1). Such method was proven to be useful for POS-tagging by Ling (Ling, 2015).

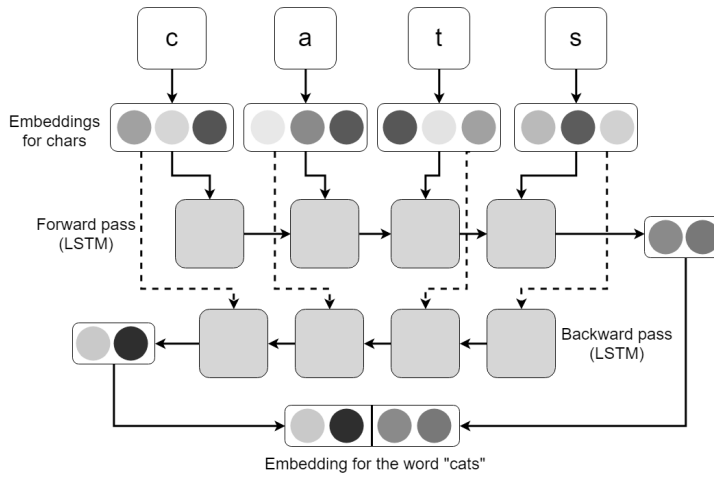


Fig. 1. The BiLSTM variant of the character-level word representation

We propose our own variant of leveraging the letters' sequence information. BiLSTM works much slower than a simple lookup in the embedding matrix. In order to increase the speed of computation, we are using a feedforward model (Char FF): two dense layers are applied over the concatenation of character embeddings (Fig. 3). Such representation can be computed much faster than the previous variant, though obviously slower than word embeddings. Moreover, the experiments showed that this representation can be trained much faster (in terms of the number of epochs) and with a smaller amount of the training data.

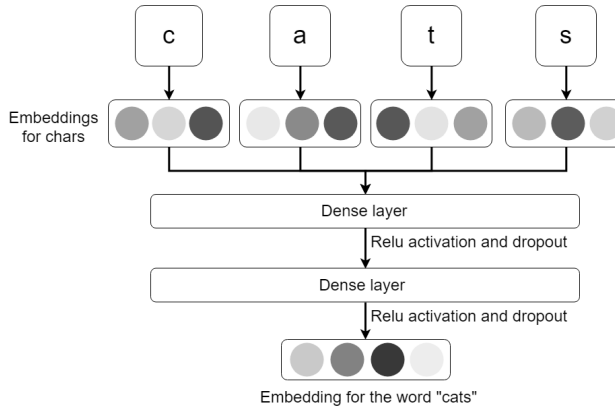


Fig. 2. The feedforward variant (Char FF) of the character-level word representation

The apparent disadvantage of such layer is the necessity to represent a word as a fixed length sequence. However, in our experiments, we found out that 98% of all words in the train set have a length less or equal to 11. Therefore, we added zero-padding in front of all words that are shorter than 11 symbols and cut the head of all words which are longer. Such trade-off between the speed and quality of the model and the loss of information encoded in the longer words seems reasonable to us.

3.3. Word Representations Pretraining

The character-level representations have much fewer parameters than the word embeddings matrices so they can be trained from scratch with the whole model. However, in this case, we are likely to suffer from overfitting. To improve the representations and utilise more unlabelled data it is possible to train such representations in the same way as the embeddings in word2vec. The disadvantage of such approach is considerably slower training process.

In order to achieve a better speed of training, we propose the following method. We are aiming to predict the word index by its representation obtained by the described variants of the character-level function. Therefore, the network consists of two parts: some character-level function, e.g. BiLSTM or the feedforward model, and the output layer with softmax activation, which predicts the word index. To ensure that

the word vector predicted by such layer is meaningful we initialize the output layer by pretrained word embeddings. We used 300-dimensional GloVe vectors for this task so we added an additional dense layer, which mapped the word vector from the character-level function space to the 300-dimensional space (Fig. 3).

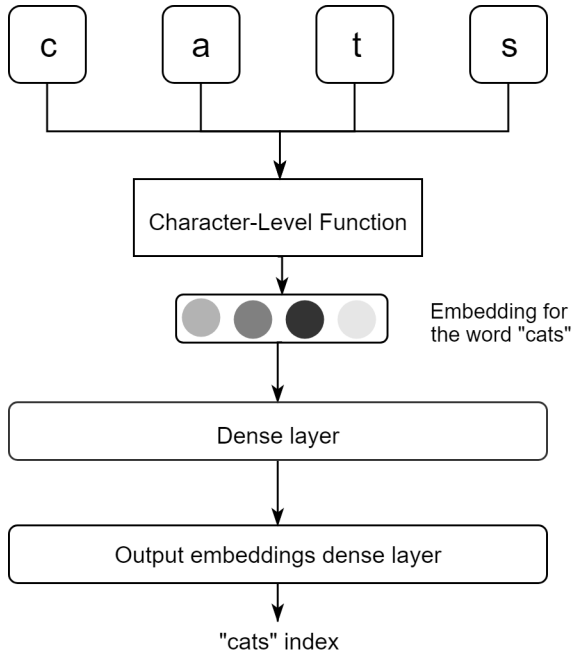


Fig. 3. Pretraining of the character-level functions

Mathematically the process can be seen as an optimization of the function F from word characters $\overline{w} = w_1 w_2 \dots w_n$ to 300-dimensional space:

$$\frac{\exp\left(w_i^T F(\overline{w}_j)\right)}{\sum_k \exp\left(w_k^T F(\overline{w}_j)\right)} \to \begin{cases} 1, i = j \\ 0, i \neq j \end{cases}$$

As a result, the representation of the word obtained by the function F should be similar by cosine distance to the appropriate GloVe-vector and less similar to all other vectors.

During the designed pretraining process, we have to iterate through all words in the embeddings' vocabulary and try to minimize cross-entropy loss. To speed up this process, we used only the words from the train set. However, it is likely that optimization over all words in the pretrained word embeddings should make the obtained representation even more robust.

It is also possible to add such auxiliary loss to the model optimization process. In this case, the character-level function is optimized using two objectives: the main, task-specific objective and our auxiliary objective. The main objective forces the character-level function to produce words' representations that are suitable to the task.

The auxiliary objective makes model learn representations similar to pretrained word embeddings. This should reduce model's overfitting and improve the quality of obtained representations.

Still, our experiments showed that the pretraining works better. We expect that in the case of auxiliary loss the model pays too much attention to the frequent words and ignores the infrequent ones.

3.4. Grammemes Embeddings

Apparently, solitary word form cannot be a good evidence for word's semantic or syntactic value: small orthographic differences may lead to completely different meanings, such as "land" vs "laud" or "taxes" vs "takes". To enhance the word's representation we propose to use grammemes embeddings.

We can represent every word with a vector where each position relates to one specific grammeme. We fill it with the estimated probability of the word to have the corresponding grammeme. For instance, the frequency of the verb "cut" is about $8.75 \cdot 10^{-5}$ its noun form has frequency $2.84 \cdot 10^{-5}$. Then we estimate the probability of the grammeme NOUN by the frequency of the noun form divided by the sum frequency of all forms of this word:

$$\frac{2.84 \cdot 10^{-5}}{2.84 \cdot 10^{-5} + 8.75 \cdot 10^{-5}} \approx 0.26$$

in this particular case.

We apply an additional linear layer with a non-linear activation on top of this vector in order to obtain not just a set of grammemes' probabilities but some interactions between them (Fig. 4). The result is similar to the feature set defined for the common linear classifiers: features as "verb + past tense" or "noun + plural" can easily be encoded by the matrix of the linear transformation. Our approach, on the other hand, gives an ability to learn some less obvious interactions between the grammemes.

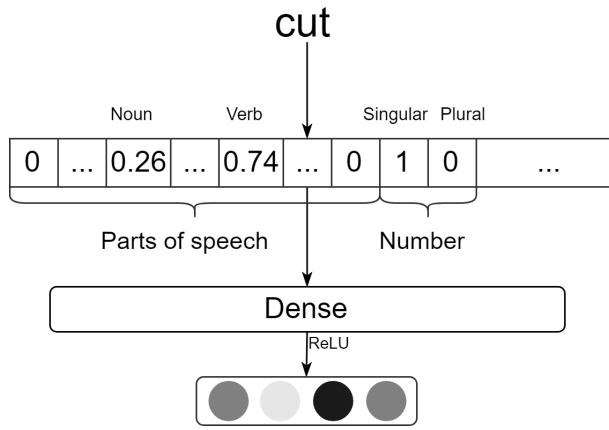


Fig. 4. The grammemes embedding example

We used ABBYY Compreno Morphology module to obtain the morphological analysis. It contains comprehensive dictionaries with grammatical value information and frequencies of words calculated on a large corpus.

4. Part-of-Speech Tagging Model Enhancements

4.1. Auxiliary Word Language Model

To make the training process more robust Rei (Rei, 2017) proposed to use the language modelling objective. The BiLSTM in his setup outputted hidden representations for each word, which were used not only to predict the word's tag but also its neighbours—the previous and the next words in the sentence.

However, a direct application of the sequence labelling model is prohibited because it would have an access to the full context in this case. The ordinary language model predicts the word using its left (or right) context only. To obtain similar behaviour of the sequence labelling model it's possible to use the forward LSTM only to predict the next word and backward LSTM to predict the previous one. Their hidden states then have to be concatenated to receive the whole information about both contexts, which would be used to predict the POS tag.

Such design helps to handle more general syntactic and semantic patterns in the data. Another advantage of this approach is in the clear ability to pretrain the model on a large number of unlabelled data. However, to our best knowledge, there was not any successful attempt to improve quality of any model using such pretraining process.

4.2. Auxiliary POS Language Model

As an alternative to the word language model, we propose to use part-of-speech language model. In our variant, we are aiming to predict the previous and the next tags using the LSTMs as well as the tag of the word.

This approach is better than the previous one in terms of training speed and required memory, because the output layer for such language model contains only tens or hundreds of elements, while in case of word language model we should consider at least ten thousand different words.

Moreover, prediction of the labels in the surrounding context forces the model to be more aware of the connections between the labels. For example, the model should learn the connection between tags "Noun" and "Adjective" or "Adverb" and "Verb".

This idea seems to be a simpler variant of structured prediction models, however, the experiments showed that its application makes a model overfit less than the model with a CRF output layer.

Nevertheless, it should be noted that it is not possible to use an unlabelled data in such variant.

4.3. Part-of-Speech Tagging with Transfer Learning

One of the most common ways to improve quality in the computer vision tasks is to adapt a model trained on a large dataset (usually, ImageNet) for classification task with a much smaller number of available data. Usually, the first layers of a pretrained model are frozen and only new output dense layers are trained on the task-specific data. It was shown that the first (frozen) layers are used mostly to extract useful features from the image, so they can be seen as invariant to the task.

We propose a similar idea for the POS-tagging. There are many datasets for English and Russian with different tagsets. Due to the differences in tagsets, we cannot apply a model trained on one dataset to another. Therefore, we should train a new model on the new dataset. However, we can just change the output layer and keep all other layers and their weights. Therefore, we obtain a pretrained model with meaningful weights.

To train this new model, we propose to freeze old layers during the first 5 epochs of training on the new dataset and tune only output layer. After we obtain good weights for the output layer, we can fine-tune all layers to get a better representation of the specific new corpus.

5. Comparison of the methods

We applied the same configuration for all the tested models and for all datasets. We used 2-layer BiLSTM with 128 units and 0.3 dropout rate. The input features were passed through the projection layer, which outputted 200-dimensional vector. As a result, all presented models had fixed size of BiLSTM layers despite the differences in the input features size. The output of the BiLSTM was projected on a lower-dimensional space using a linear layer with Batch Normalization (Ioffe, 2015) followed by the output layer with the number of units equals to the number of predicted classes.

Fig. 5 shows the final variant of the model. This variant has grammemes and character-level embeddings, CRF output layer and it was trained using POS language model auxiliary loss. The parameters of layers are specified in parentheses.

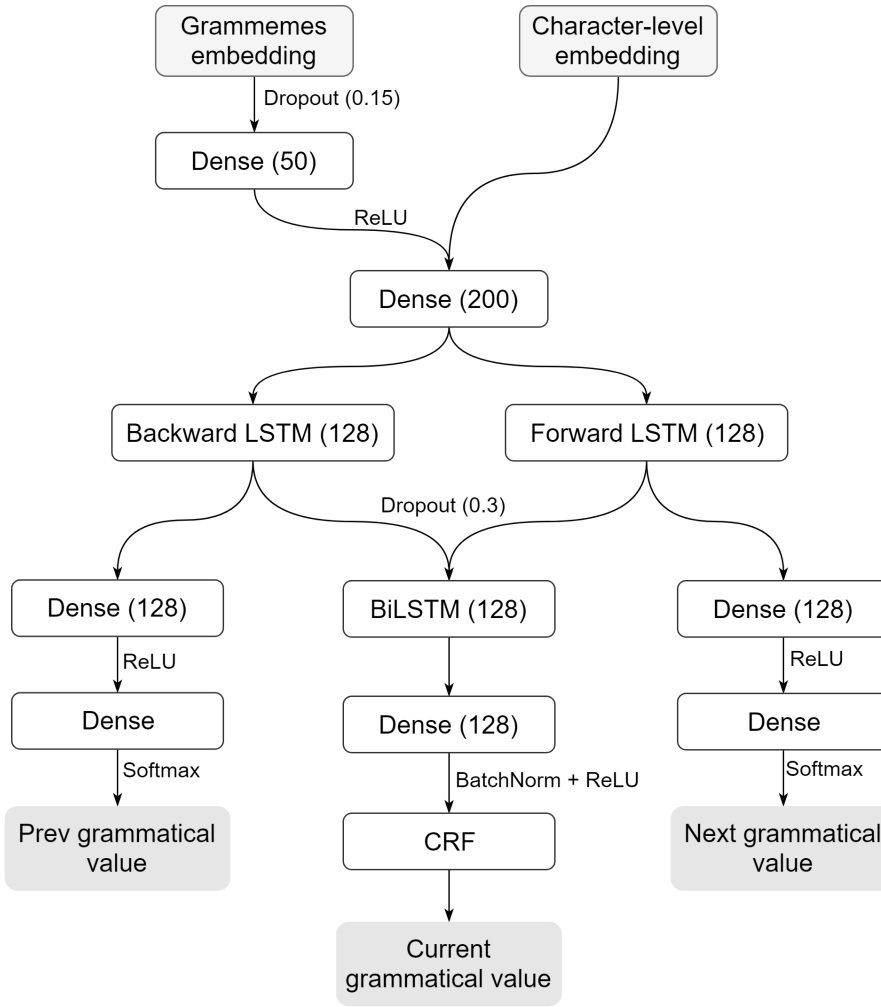


Fig. 5. The final model architecture.

In the next sections we are going to explore the contribution of each distinct component of the final model. We checked the quality on the Penn Treebank dataset for English language and Syntagrus (from Universal dependences 2.2) and Gikrya (from MorphoRuEval-2017) datasets for Russian. Their number of tokens are presented in Table 1.

Table 1. Number of tokens in used corpora

Dataset	Train	Development	Test
PTB	912,344	131,768	129,654
Syntagrus	871,082	118,630	117,470
Gikrya	977,567	108,581	19,560

Penn Treebank tagset contains just 45 different tags (including punctuation tags). This corpus is somewhat standard for POS taggers evaluation—most of the well-known models were trained on it.

The Gikrya dataset has 304 grammatical values. However, during Morpho-RuEval-2017 the quality of the models was evaluated on the subset of these grammatical values (e.g., animacy category was not included into this subset). We trained our model to predict the whole set of tags but the evaluation on the test set was performed only for the specified subset (about 250 distinct grammatical values).

The Syntagrus contains 908 different grammatical values. This corpus seems better than Gikrya in terms of test set size, but we do not know any clearly state-of-the-art result shown on this corpus—which is why we evaluated our models on both corpora.

5.1. Word Representations

5.1.1. Different Character-Level Functions

Firstly, we compared the quality of the proposed feedforward character-level function (Char FF) with more classic BiLSTM one (Char BiLSTM).

In our experiments, the Char FF model consisted of two linear layers with ReLU activations and small dropouts. The linear layers contained 500 and 200 units respectively, the dropout rate was equal to 0.15. The BiLSTM model contained 150 units. The size of character embeddings was 24.

Table 2. Comparison of BiLSTM and feedforward character-level models on development and test sets

Dataset	Char BiLSTM	Char FF
Syntagrus	**95.23%** / **95.39%**	94.98% / 95.16%
Gikrya	96.48% / **94.69%**	**96.68%** / 94.63%
PTB	97.02% / 96.98%	**97.32%** / **97.26%**

These two variants of the character-level function performed roughly similar (Table 2) but the proposed feedforward model converged in a smaller number of epochs and worked much faster than BiLSTM.

5.1.2. Effect of Pretraining

It seems quite reasonable to expect that the pretraining of character-level representation using the word vectors should increase the performance of the model. The experiments showed that the model with pretrained character-level function (Char FF Pretrained) has higher quality during the first few epochs and achieves accuracy about 0.1–0.2% greater than the model without pretraining (Table 3).

Table 3. Comparison of models with and without pretrained character-level representations

Dataset	Char FF	Char FF (Pretrained)
Syntagrus	94.98% / 95.16%	**95.22%** / **95.36%**
Gikrya	96.68% / **94.63%**	**96.88%** / **94.63%**
PTB	97.32% / 97.26%	**97.40%** / **97.31%**

Such improvement leads to 4–5% error rate reduction (ERR) on Russian datasets and 2–3% ERR on PTB, which seems significant enough given the simplicity and cheapness of the pretraining process.

5.1.3. Grammemes Embeddings

The models with grammemes embeddings converge much faster. Grammemes embeddings seriously improved the models' performance on both Russian datasets. However, the English model without them achieved approximately the same quality after a large number of epochs (on the other hand, the model with grammemes embeddings needed roughly twice as lesser epochs to converge).

Table 4. Comparison of models with and without grammemes embeddings

Dataset	Char FF (Pretrained)	+ Grammemes
Syntagrus	95.22% / 95.36%	**96.77%** / **97.00%**
Gikrya	96.88% / 94.63%	**98.07%** / **95.36%**
PTB	97.40% / **97.31%**	**97.43%** / 97.30%

Obtained results may be explained by the differences in Russian and English morphologies. Russian language has considerably more complicated morphological system. As a result, Russian grammemes embeddings are far more informative. Moreover, it is usually harder to predict morphological tag for Russian words using merely word's form.

5.2. Language Model Auxiliary Objectives

Both word and POS auxiliary objectives effectively work as regularizer: a model with them overfits considerably slower and achieves better results. Slower overfitting in our case means that the difference between train and development accuracies remained insignificant even after 200 epochs, while without such auxiliary objective this difference became more than 1% after the first 100 epochs on all datasets.

The achieved results are presented in Table 5. Word and POS LM objectives gave similar results on the Penn Treebank. However, POS LM has shown clearly better performance on the Russian datasets.

Table 5. Comparison of models with different auxiliary objectives

Dataset	Char FF (Pretrained) + Grammemes	+ Word LM	+ POS LM
Syntagrus	96.77% / 97.00%	96.69% / 96.96%	**96.97% / 97.24%**
Gikrya	98.07% / 94.85%	97.91% / 96.30%	**98.12% / 96.72%**
PTB	97.43% / 97.30%	97.57% / 97.49%	**97.57% / 97.49%**

The most significant improvement was achieved on the Gikrya test set. Yet, the 7% ERR on the PTB test set and 8% ERR on Syntagrus test seem good enough to consider the proposed auxiliary objective successful.

5.3. CRF Layer

Usage of CRF output layer usually leads to noticeable improvement in the model's quality. However, we could not achieve better results with the CRF layer. We expect that the main reason is our POS LM objective: it forces models to learn the same connections between the tags as CRF layer does.

Table 6. Comparison of models with and without CRF layer

Dataset	Char FF (Pretrained) + Grammemes + POS LM	+ CRF
Syntagrus	**96.97% / 97.24%**	96.72% / 96.97%
Gikrya	**98.12% / 96.72%**	98.07% / 96.65%
PTB	97.57% / 97.49%	**97.60% / 97.51%**

5.4. Transfer Learning

Finally, we applied the proposed transfer learning process. We pretrained two models. The first one was trained on Compreno corpus with about 10 million tokens. The tagset contains 1040 different grammatical values. The second one was trained on the Gikrya corpus. The results are presented in Table 7.

Table 7. Qualities of pretrained models on Syntagrus corpus

Model	Accuracy
Base	96.97% / 97.24%
Compreno pretrained	98.18% / 98.29%
Gikrya pretrained	**98.21% / 98.33%**

Gikrya and Syntagrus have similar tagset (based on Universal Dependencies standard) while Compreno's tagset and applied conventions are very different from

Syntagrus corpus. Therefore, it is understandable that Gikrya pretrained model achieved better quality.

As a result, we achieved very significant improvement using quite a simple process.

5.5. Summary

To conclude, we applied few tricks to achieve results that are on par with state-of-the-art results on considered datasets. We did not try to optimize the hyper-parameters of the models so the work should be seen mostly as a proof-of-concept.

Table 8. Comparison with the existing models on PTB dataset

Tagger	Test Acc
Manning (2011)	97,32%
Søgaard (2011)	97,50%
Santos (2014)	97,32%
Ling (2015)	**97,78%**
Ma (2016)	97,55%
Choi (2016)	97,64%
Rei (2017)	97,43%
This work	97,51%

Table 9. Comparison with results on MorphoRuEval-2017 on Gikrya dataset

	Modern literature	News	VKontakte
Best closed track model	94.16%	93.71%	92.29%
Best open track model	**97.45%**	97.37%	**96.52%**
Our model	96.46%	**97.97%**	95.64%

The final model is worse than the best PTB models. On the other hand, it does not use word embeddings. That means that our model is much smaller. To our best knowledge, the achieved result is the best for models without word embeddings.

The model also shows poorer performance than the best model on MorphoRuEval-2017. However, the best model was trained using additional data and it used word embeddings too.

6. Conclusions

We proposed a new method of word's representation using character-level feed-forward neural network and a way to pretrain such representations with existing word embeddings. This variant of words representations seems to perform better than the previous ones. Moreover, the pretraining process helps to use the information encoded in the word embeddings implicitly. That means we can expect syntactic and semantic meaningfulness of the representations, which can be found in the word2vec vectors.

We described a way to encode an additional information about the grammatical value of the word in the grammemes embedding. Such embeddings are useful as a way to regularise the word representation and provide it with an additional morphological information.

We also proposed a novel approach to multi-task learning for sequence labelling tasks and tested it on POS tagging problem. The achieved results a comparable to the state-of-the-art in this area.

Finally, we proved the possibility of transfer learning for the POS tagging task. We achieved almost 40% ERR on Syntagrus dataset by applying this process.

Summing up, the proposed tricks are aimed to improve the quality of input features and to regularize the learned objective. We consider them simple enough to be applicable for most of NLP tasks and the achieved results seem reasonably good.

References

1. *Anastasyev D. G. et al.* (2017) Part-of-speech tagging with rich language description, Computational linguistics and intellectual technologies: Proceedings of the International Conference "Dialog 2017". Vol. 1, pp. 2–13
2. *Anisimovich K. V., et al.* (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
3. *Choi J.* (2016), Dynamic Feature Induction: The Last Gist to the State-of-the-Art, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (NAACL 2016), San Diego, CA, pp. 271–281.
4. *Hochreiter S., Schmidhuber J.* (1997), Long Short-Term Memory, Neural Computation, Vol. 9, Issue 8, pp 1735–1780.
5. *Huang Z., Xu W., Yu K.* Bidirectional LSTM-CRF models for sequence tagging, arXiv:1508.01991.
6. *Ioffe S., Szegedy Ch.* (2015), Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, arXiv:1502.03167
7. *Korobov M.* (2015), Morphological Analyzer and Generator for Russian and Ukrainian Languages, Analysis of Images, Social Networks and Texts, Vol. 542, pp. 320–332.
8. *Ling W. et al* (2015), Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation, available at: https://arxiv.org/abs/1508.02096.
9. *Ma X., Hovy Ed.* (2016), End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF, arXiv:1603.01354.
10. *Rei M.* (2017), Semi-supervised Multitask Learning for Sequence Labeling, arXiv.org:1704.07156.
11. *Søgaard A.* (2011), Semisupervised condensed nearest neighbor for part-of-speech tagging, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp. 48–52
12. *Sorokin A., et al.* (2017) MorphoRuEval-2017: an Evaluation Track for the Automatic Morphological Analysis Methods for Russian, Proceedings of the International Conference "Dialog 2017". Vol. 1, pp. 297–313

DISCOVERING DIALECTAL DIFFERENCES BASED ON ORAL CORPORA[1]

Andriyanets V. (blindedbysunshine@gmail.com),
Daniel M. (misha.daniel@gmail.com)

International Linguistic Convergence Laboratory, NRU HSE, Moscow, Russia

Pakendor B. (brigitte.pakendorf@cnrs.fr)

Laboratoire "Dynamique du Langage", UMR5596, CNRS & Université de Lyon, Lyon, France

This paper discusses a method to detect statistically significant linguistic differences between corpora while factoring in possible variability within the very corpora to be compared. Specifically, we compare two small corpora of dialects of Even, Bystraja and Lamunkhin Even, in an attempt to identify morphemes that are more frequent in either of the corpora. To investigate whether this difference might be due to an over-representation of a speaker who happens to be an outlier in terms of using a particular morpheme, we use DP, a measurement of evenness of the distribution of a specific linguistic feature across subcorpora of the same corpus.

Keywords: linguistic corpora, Even, dialect variation, outliers

ИЗВЛЕЧЕНИЕ ДИАЛЕКТНЫХ РАЗЛИЧИЙ НА МАТЕРИАЛЕ УСТНЫХ КОРПУСОВ

Андриянец В. (blindedbysunshine@gmail.com),
Даниэль М. (misha.daniel@gmail.com)

Международная лаборатория языковой конвергенции, НИУ ВШЭ, Москва, Россия

Пакендорф Б. (brigitte.pakendorf@cnrs.fr)

Лаборатория «Динамика языка», UMR5596, Национальный центр научных исследований и Университет Лион, Лион, Франция

[1] This article was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy by the Russian Academic Excellence Project '5-100'. BP is grateful to the LABEX ASLAN (ANR-10-LABX-0081) of Université de Lyon for its financial support within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) of the French government operated by the National Research Agency (ANR).

1. Introduction

Even is a North Tungusic language spoken in Siberia and the Russian Far East. The overall number of speakers is probably not more than two to three thousand, and these are settled over a vast territory in small individual speech communities. It is a dialectally diverse language, with 13 dialects (диалекты) and up to 24 subdialects (говоры) recognized [Burykin 2004: 85]. This paper compares data from two Even dialects, Lamunkhin Even and Bystraja Even. Lamunkhin is the westernmost still viable dialect of Even spoken in the village Sebjan-Küöl in Yakutia. It has been in close contact with the Turkic language Sakha (Yakut) for decades, and Sakha influence is registered at all levels of the language [Pakendorf 2009]. Bystraja, one of the easternmost Even dialects, is spoken in the Bystraja district in central Kamchatka. The extent to which it may have undergone contact influence from the neighbouring language Koryak is yet to be elucidated. This dialect is undergoing a shift to Russian, with no confident speakers younger than 40–45 years of age.

The texts in the corpora were collected in several field trips between 2007 and 2015 by Brigitte Pakendorf with contributions by Natalia Aralova. These are spoken texts, mostly narratives that were glossed and translated in Field Linguist's Toolbox[2], with the majority time-aligned in ELAN[3]. The Lamunkhin corpus comprises ~50,000 tokens and is recorded from 37 speakers, and the Bystraja corpus comprises ~34,000 tokens and is recorded from 26 speakers. Importantly, as will be described in Section 2, neither corpus is balanced in terms of contributions by individual speakers.

Even dialects are known to differ across linguistic domains [cf. Rišes & Cincius 1952], not only in lexicon, but also in phonology, morphology and syntax. The two dialects included in this study are no exception (e.g. [Matić & Pakendorf 2013], [Pakendorf & Krivoshapkina 2014], Pakendorf to appear). For instance, the simultaneous converb –nikEn[4] is used more frequently in the Lamunkhin dialect, where in addition the converb of 'say' has taken on extended functions, such as that of complementizer. Furthermore, habitual aspect is expressed with the 'generic' suffix in the Bystraja dialect and with the 'habitual' suffix in the Lamunkhin dialect[5] (1). In the nominal domain, the dative case is extending to addressees of verbs of speech in the Lamunkhin dialect, a function that is fulfilled only by the allative case in the Bystraja dialect (2).

---

2 https://software.sil.org/toolbox/

3 https://tla.mpi.nl/tools/tla-tools/elan/

4 Even retains vestiges of vowel harmony, so that vowels in suffixes can vary between [e] and [a], and consonants undergo various assimilation processes. When suffixes are shown in isolation, capital letters indicate phonemes that undergo changes.

5 Abbreviations used in the glosses are: agnr: agent nominalizer; all: allative; ant: anterior; cvb: converb; dat: dative; dim: diminutive; dist: distal (demonstrative); dp: discourse particle; fut: future; gnr: generic; hab: habitual; nfut: non-future; pf: perfect; poss: possessive; prog: progressive; ptc: participle; qual: qualitative; sim: simultaneous; Y: Sakha (Yakut) borrowing

(1a) Bystraja (RME_fox_wolf_053)
nan *tačin* *tar* ***go:-niken*** *ereger* *ŋene-d-**ʤo:t**-te-n*
and dist.qual dist **say-sim.cvb** always go-prog-**gnr**-nfut-3sg
'Saying like this [the fox] was coming all the time…'

(1b) Lamunkhin (stado#10_SEN_poems_084)
Mitja *ihu-riʤi* *ǫralči-mŋa* *bi-ʤi-n* ***go:-niken*** *ʤǫmkak-**kara**-m*
Mitja grow-ant.cvb herd.reindeer-agnr be-fut-3sg **say-sim.cvb** think-**hab**[nfut]-1sg
'I think that Mitja, having grown up, will be a reindeer herder.'

(2a) Bystraja (NAT_rabotajushaja_010)
ose:l-če-l *eniŋe:-wu* *gia-**tki*** *atikan-**taki***
go:n-ʤid-de-n
become.tired-pf.ptc-pl grandmother-poss.1sg next-**all** old.woman-**all**
say-prog-nfut-3sg
'My tired grandmother said to the other old woman,…'

(2b) Lamunkhin (KKK_Emcheni_056)
asatka-čan *bǫllaɣina* *tar* *ǫmǫlgǫ* *kuŋa-**du*** *go:n-če*
girl-dim dp.Y dist boy child-**dat** say-pf.ptc
'And the girl said to the boy:…'

Such differences may result from divergence of dialects of the same language (i.e. independent evolution of mutually isolated linguistic systems) or convergence of dialects or languages (via contact-induced changes in socially and geographically adjacent linguistic systems). However, as yet no comprehensive account of the morphosyntactic differences between the two dialects exists. In this paper we undertake the first step towards filling this gap by assessing the issues involved in detecting significant quantitative differences between the dialects using small corpora of oral narratives. To the best of our knowledge, this is the first attempt at using text corpora to evaluate differences between dialects. We assume that any statistically significant quantitative difference deserves a qualitative interpretation, but such interpretation is outside the scope of this paper.

2. Data

The texts in the two corpora were glossed in Toolbox over a period of several years, and contained, in addition to typos, some inconsistencies and traces of the evolution of the researcher's conceptions that needed to be corrected in an initial step. Although most of the texts are narratives, in some of them speakers other than the narrator step in for a phrase or two, and a few are conversations. Each annotation unit (roughly, a sentence, to the extent that the notion of sentence can be applied to spoken text) is thus associated with a speaker (represented by a two- or three-letter code).

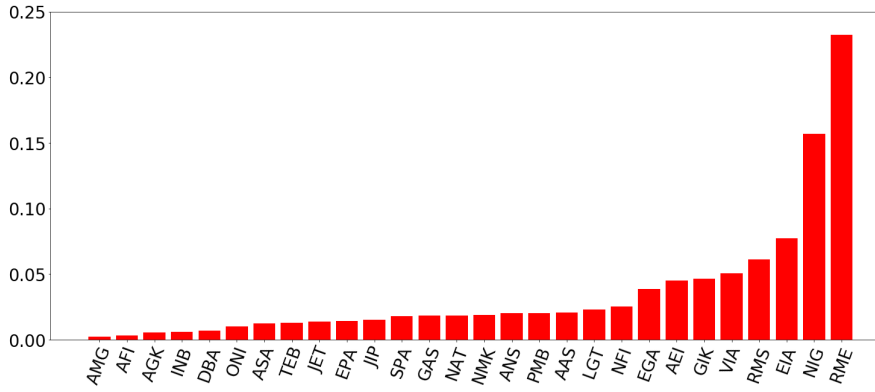


Fig. 1. Relative contributions of individual speakers to the Bystraja corpus

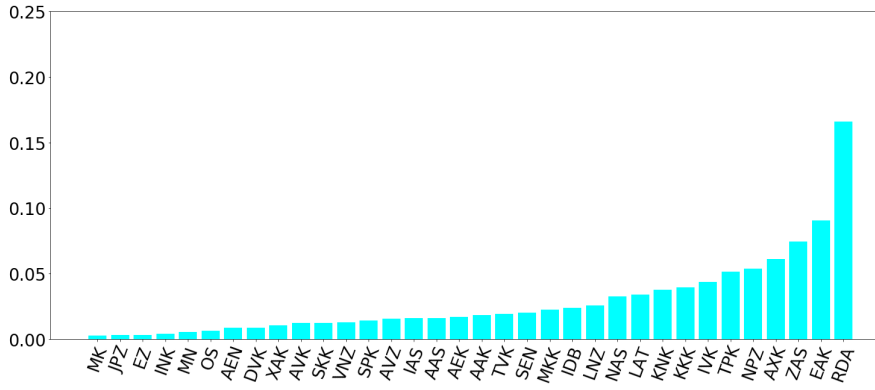


Fig. 2. Relative contributions of individual speakers to the Lamunkhin corpus

As discussed above, the Bystraja corpus contains 34,000 tokens and the Lamunkhin corpus contains 50,000 tokens. However, both corpora are very unbalanced in terms of contributions of individual speakers (Fig. 1, 2). This is particularly striking in the Bystraja corpus, where the two speakers with the largest contributions (RME and NIG) account for 39% of the corpus (with just RME's contribution amounting to 23%)—practically equal to the contribution of the ten speakers with the next largest contributions (41%). Although the Lamunkhin corpus is somewhat more balanced, even here one third was contributed by only three speakers (RDA, EAK, and ZAS), with RDA alone accounting for 13%. Estimates of dialectal frequency differences based on such unbalanced corpora might well be distorted by the linguistic idiosyncracies of the speakers who contributed the most, rather than reflecting actual differences between the dialects.

3. Method

Notwithstanding the heterogenous nature of the corpora, our first approach to a quantitative analysis of the data was the conventional log-likelihood measure (as used in [Rayson & Garside, 2000]) to compare the relative frequencies of each suffix in the corpora. The log-likelihood comparison is based on the overall size of the compared corpora and the frequency of a given suffix in each corpus[6]. The comparison indicates noticeable differences between the corpora in the use of several morphemes; here, we focus on the spatial case markers (locative, allative and dative, which is also used as a directional marker) and converbs[7]. Table 1 shows the observed and the expected (under the assumption of an even distribution in the two corpora) frequencies of the markers in the two dialects.

The difference in the frequencies is overwhelming in the case of the simultaneity converb and the allative, but it is also deemed significant for the other suffixes (log-likelihood values above 10 are considered to be statistically significant).

A potential problem with this analysis is that, as explained in Section 2 above, the two corpora are unbalanced in terms of speaker representation. Given the small size of the corpora (~50,000 tokens for Lamunkhin and ~34,000 tokens for Bystraja) and the large variation in speaker contributions illustrated in Fig. 1 and Fig. 2, the differences between the corpora shown in Table 1 might reflect not actual differences between the dialects, but rather be the result of a greater contribution to either of the corpora by a speaker or speakers who are (in)frequent users of a particular grammatical category.

Table 1. Log-likelihood comparison for spatial case markers and converb suffixes, ordered by decreasing log-likelihood (LL) value

Suffix	Observed, Lamunkhin	Observed, Bystraja	Expected, Lamunkhin	Expected, Bystraja	LL
Simultaneity converb -nikEn	1,083	59	688.72	453.28	739.84
Allative case -t(E)ki	78	378	275.00	181.00	360.17
Multiplicative converb ntEkEn	101	0	60.91	40.09	102.15

[6] The expected value of a morpheme for each corpus is calculated on the basis of the overall frequency of the morpheme in both corpora in relation to the relative size of each corpus. See http://ucrel.lancs.ac.uk/llwizard.html for a technical description of the log-likelihood calculation.

[7] Converbs are subordinate verb forms typically introducing adverbial clauses and specifying a temporal, logical or other relation of the event described by such subordinate clauses to the main clause. See [Haspelmath, König 1995] for a typological overview of the category. They are not obligatory, since they can be replaced by alternative subordinate constructions using case-marked participles or by chains of finite verb clauses.

[8] DS = different-subject, i.e.subordinate and main clause subjects are non-coreferential

[9] SS = same-subject, i.e. subordinate and main clause subjects are coreferential

Suffix	Observed, Lamunkhin	Observed, Bystraja	Expected, Lamunkhin	Expected, Bystraja	LL
Anteriority converb -Riʤi	721	237	577.75	380.24	95.30
Purposive converb -DE	307	405	429.40	282.60	85.46
Dative case -Du	699	299	601.88	396.12	40.93
DS[8] conditional converb -REk	488	197	413.11	271.88	35.66
Locative case -(du) LE	966	503	885.93	583.07	18.56
SS[9] conditional converb -mi	442	223	401.05	263.95	10.75

To assess the unevenness of the distribution of different morphemes across speakers in each corpus we used the DP metric [Gries 2008]. DP is a measurement that shows how evenly a feature is distributed across corpus parts. Importantly, DP is based on the cumulutive difference between the expected and observed numbers of uses in each subcorpus against the total number of its uses in the whole corpus, rather than on the frequency of a category in the subcorpus, as log-likelihood is. Absolute values of pairwise differences are added and divided by two, and the resulting DP value lies between 0 (absolutely even distribution) and 1 (infinitely uneven distribution)[10].

For the Even data, we calculated the evenness of the distribution of a morpheme between the speakers of each dialect separately for each dialect using an in-house Python script. For each of the two corpora, the calculation showed that all frequently occurring suffixes and, importantly, most of the suffixes we were primarily interested in (i.e. the locative and dative case marker as well as converbs) had DP values lower than 0.2. For instance, the DP value for the conditional converb *-REk* is ~0.19 in the Bystraja corpus and ~0.17 in Lamunkhin; that for the locative *-(du)LE* is ~0.09 (Table 2). We interpret this as an indication that the feature is more or less evenly distributed. More precisely, it means that less than 20 percent of the distribution of a suffix is not where it is expected to be under the assumption of its even distribution between the speakers.

The following table shows the DP values for the suffixes of interest in this study, computed across all speakers in each corpus and compared to the corresponding log-likelihood values. Note that we excluded the multiplicative converb shown in Table 1 because it does not occur in the Bystraja dialect at all, and therefore the DP metric cannot be calculated for this dialect.

[10] DP = sum(abs(exp-obs))/2.

Table 2: DP values within corpora for the spatial case markers and converb suffixes compared to log-likelihood values[11]

Suffix	DP Lamunkhin	DP Bystraja	Log-likelihood
locative (nouns) -(du)LE	0.085	0.098	18.56
DS conditional converb -REk	0.165	0.187	35.66
dative (nouns) -Du	0.194	0.168	40.93
purposive converb -DE	0.208	0.155	85.46
SS conditional converb -mi	0.248	0.168	10.75
simultaneity converb -nikEn	0.139	0.294	739.84
anteriority converb -Riʤi	0.149	0.318	95.30
allative (nouns) -t(E)ki	0.360	0.170	360.17

From the data in Table 2 it appears as if the simultaneity converb –nikEn and anteriority converb—Riʤi, as well as the allative case suffix –t(E)ki are quite unevenly distributed in the Bystraja and Lamunkhin corpus, respectively, since the DP values for these suffixes are three times higher than the DP value calculated for the locative case marker (which is similar to the lowest DP value calculated for a sample of common English function words in the BNC Sampler corpus; [Gries 2008: 421]). This therefore indicates that the putative dialectal differentiation emerging in the log-likelihood scores (Table 1) needs to be evaluated with caution, since the observed differences might be due to speaker idiosyncracies rather than to true dialectal differences.

4. Discussion and Conclusion[12]

Dialectal differences between languages may be due to both linguistic divergence (independent innovations such as the loss of inherited categories) and linguistic convergence (innovations due to the influence of other dialects or unrelated languages). Often, descriptions of dialectal differentiation are based on categorical differences between varieties (presence or absence of a category). Categorial changes can, however, be preceded by changes in usage [Johanson 1999: 52]; [Aikhenvald 2002: 238]; [Heine & Kuteva 2005: 50]; frequency of use therefore needs to be taken into account in evaluations of dialect divergence. However, for any qualitative interpretation of the frequency differences between two linguistic varieties, it is important to ensure that the observed differences reflect true dialectal divergence rather than biases in the data under comparison. This is especially important when working with minority languages where the data are scarce and one cannot from the outset exclude

[11] The DP values are arranged by increasing order of the higher of the two DP values for each morpheme.

[12] We sincerely thank an anonymous reviewer for her/his thoughtful feedback on our manuscript, which has greatly influenced our interpretation of our results.

utterances by under- or overrepresented speakers in order to homogenize potentially heterogenous corpora.

It is therefore important when evaluating dialectal differences based on oral corpora of minority languages to first apply methods that take into account potential heterogeneity in speaker contributions. In this paper, we found skewed distributions in the frequency of use of spatial case markers (especially the allative) and of converb suffixes (especially the simultaneous converb) between two small dialectal corpora, as reflected in their very high log-likelihood values (Table 1). While this may indicate dialectal differentiation, the DP metric [Gries 2008] indicates that the difference in their use varies greatly between speakers. Further tests are therefore necessary before one can safely conclude that the apparent frequency differences between the two corpora are indeed due to dialectal differentiation.

A first qualitative assessment of the impact of the heterogeneity in speaker contributions is provided by the breakdown of frequency of use by individual speakers of the three morphemes with the highest DP values: the allative case marker for the Lamunkhin dialect and the anterior and simultaneous converb for the Bystraja dialect (Fig. 3–5). All three suffixes are used less frequently than expected in the respective dialect (Table 1). Were this underuse driven by individual speakers' idiosyncracies, we would expect to find that some of the speakers with large contributions to the corpora use these suffixes less than expected.

As can be seen from the figures, of the ten speakers with the highest contribution to the Lamunkhin corpus, NPZ, and especially TPK and KNK use the allative suffix less often than would be expected merely by the number of tokens they contributed to the corpus. Of the ten speakers with the highest contribution to the Bystraja corpus, LGT, NFI, and AEI use both the simultaneity and the anterior converb less than expected judging from the size of their contributions. Although it is not the speakers with the absolutely largest contributions who appear to be underusing these morphemes, the impact of this potential speaker bias should be evaluated before a qualitative investigation of the putative dialectal differences can be approached. One way to do this would be to exclude these speakers from the corpus of the respective dialect and to redo the log-likelihood calculations to assess the degree of impact of their idiosyncratic behaviour on the differences between the corpora.

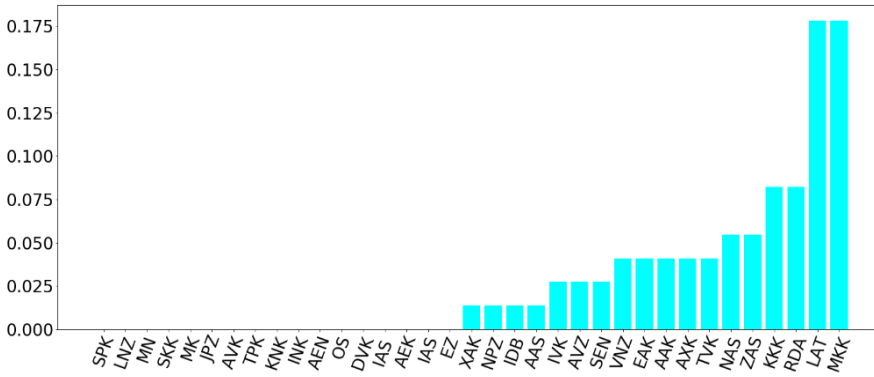


Fig. 3. Frequency of use of allative suffix by individual speakers in the Lamunkhin dialect

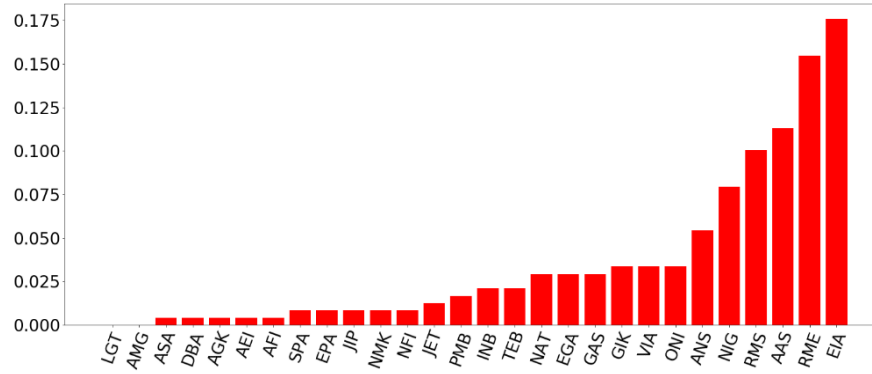


Fig. 4. Frequency of use of anteriority converb by individual speakers in the Bystraja dialect

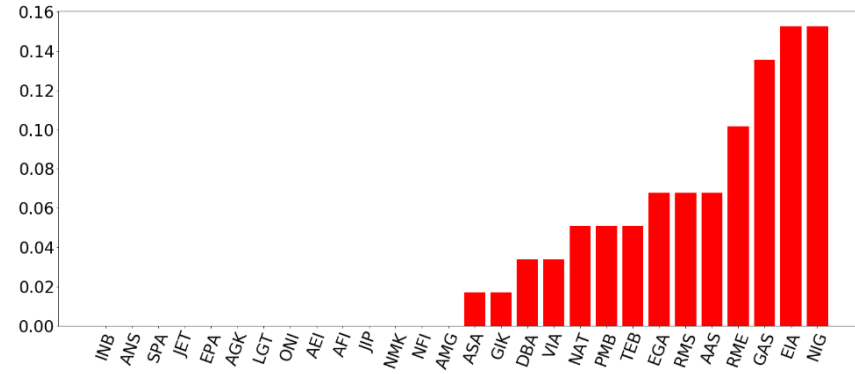


Fig. 5. Frequency of use of simultaneity converb by individual speakers in the Bystraja dialect

The DP metric thus provides a useful means to highlight potential heterogeneities in corpus data that need to be investigated before conclusions can be drawn concerning differences between corpora. However, it is not evident how to interpret the DP values and which values are to be considered high. The DP statistic is relatively sensitive to small numbers [Gries 2008: 423], so that DP values for relatively small corpora, such as those resulting from linguistic documentation projects, are expected to be relatively large—but it is not clear how large, and it is therefore hard to judge which values would be within the normal range for a given corpus. A potential solution to this problem would be to follow the heuristic suggested by [Gries 2008: 423], namely to "evaluate distributional statistics (…) in terms of the ranking of words in comparison to other words rather than their absolute values in isolation". In the corpora studied here, none of the DP values estimated for the morphemes of interest is among the highest ranking values for the overall set of morphemes found in the corpora. For instance, in the Bystraja corpus, the DP value for the anteriority converb only ranks 33rd out of 183 morphemes, i.e. 32 morphemes in the corpus have a higher DP value. Similarly, the DP value for the allative case suffix in the Lamunkhin corpus is in rank 26 (out of a total of 203 morphemes). These observations would seem to attenuate the conclusion that the DP values of 0.31 and 0.36 for the morphemes of interest indicate their substantial underdispersion in the corpora—but this is not sufficiently clear.

It is thus clearly necessary to perform further analyses, both to better understand the performance of the DP metric in small corpora as well as to obtain a trustworthy estimate of true dialectal frequency differences.

References

1. *Aikhenvald, A. Y.* (2002), Language Contact in Amazonia. Oxford, New York: Oxford University Press.
2. *Burykin, Aleksej A.* (2004), Jazyk maločislennogo naroda v ego pis'mennoj forme. Sociolingvističeskie i sobstvenno lingvističeskie aspekty [The language of a minority people in its written form. Sociolinguistic and linguistic aspects]. St Petersburg: Peterburgskoe Vostokovedenie.
3. *Gries S. T.* (2008), Dispersions and adjusted frequencies in corpora, International journal of corpus linguistics, Vol. 13, 4, pp. 403–437.
4. *Haspelmath M., König E.* (eds.) (1995), Converbs in cross-linguistic perspective. Structure and meaning of adverbial verb forms. Berlin—New York: Mouton de Gruyter.
5. *Heine, B., Kuteva, T.* (2005), Language Contact and Grammatical Change. Cambridge: Cambridge University Press
6. *Johanson, L.* (1999), The dynamics of code-copying in language encounters. In: Brendemoen, B. et al. (eds), Language encounters across time and space. Studies in language contact. Oslo: Novus forlag: pp. 37–62.
7. *Matić, D., Pakendorf, B.* (2013), Non-canonical SAY in Siberia: Areal and genealogical patterns. Studies in Language Vol. 37, 2, pp. 356–412.

8. *Pakendorf, B.* (2009), Intensive Contact and the Copying of Paradigms: An Ėven Dialect in Contact with Sakha (Yakut). Journal of Language Contact Varia 2, pp. 85–110.
9. *Pakendorf, B.* (to appear), Expressing equality, similarity, and pretense in Even (Northern Tungusic, Siberia). To appear in : Treis, Y., Chamoureau, C. (eds) : special issue of Faits des Langues on "Comparaisons d'égalité et de similitude et expression de la simulation".
10. *Pakendorf, B., Krivoshapkina, I. V.* (2014), Ėven nominal evaluatives and the marking of definiteness, Linguistic Typology Vol. 18, 2, pp. 289–331.
11. *Rayson P., Garside R.* (2000), Comparing corpora using frequency profiling, Proceedings of the workshop on Comparing Corpora, Association for Computational Linguistics, pp. 1–6.
12. *Rišes, L. D., Cincius, V. I.* (1952), Kratkij očerk grammatiki ėvenskogo (lamutskogo) jazyka. [A short grammar sketch of the Even (Lamut) language.]. Russko-Ėvenskij Slovar', pp. 693–777. Moscow: Gosudarstvennoe izdatel'stvo inostrannyx i nacional'nyx slovarej.

ЧАЙНИК ДОЛГО (НЕ) ЗАКИПАЕТ, КОМПЬЮТЕР ДОЛГО (НЕ) ЗАГРУЖАЕТСЯ…[1]

Апресян В. Ю. (valentina.apresjan@gmail.com)[1,2],
Шмелев А. Д. (shmelev.alexei@gmail.com)[2–4]

[1]Национальный исследовательский университет Высшая школа экономики; [2]Институт русского языка им. В. В. Виноградова РАН; [3]Московский педагогический государственный университет; [4]Православный Свято-Тихоновский гуманитарный университет

RUSSIAN CONSTRUCTIONS CHAINIK DOLGO (NE) ZAKIPAET, KOMP'IUTER DOLGO (NE) ZAGRUZHAETSIA…

Apresjan V. Ju. (valentina.apresjan@gmail.com)[1,2],
Shmelev A. D. (shmelev.alexei@gmail.com)[2–4]

[1]National Research University Higher School of Economics; [2]Vinogradov Institute of Russian Language, Russian Academy of Sciences; [3]Moscow Pedagogical State University; [4]St Tikhon's Orthodox University

The paper deals with a curious phenomenon of quasi-synonymy that occurs in Russian between sentences with non-negated and negated predicates in the construction with the adverb *dolgo* 'for a long time'. Consider sentences like *Chainik dolgo zakipal* 'It took the kettle a long time to boil, lit. Kettle for a long time boiled' vs. *Chainik dolgo ne zakipal* 'It took the kettle a long time to boil, lit. Kettle for a long time not boiled'. The paper is an attempt to define the semantic and pragmatic mechanisms of such quasi-synonymy,

[1] В основу настоящей статьи положен расширенный вариант доклада, представленного на конференции «Соотношение времени и вида в типологической и контрастивной перспективе» (Лилль, апрель 2018). Корпусное исследование семантических и прагматических механизмов квазисинонимии было произведено в рамках работы по гранту РГНФ на 2016–2018 годы, 16-04-00302 «Подготовка третьего выпуска Активного словаря русского языка» (руководитель Ю. Д. Апресян). Исследование семантики и прагматики рассматриваемой конструкции на основе переводных текстов (при помощи параллельных корпусов) было произведено при поддержке РНФ (проект № 16-48-03006, «Семантический анализ переводов в сопоставительной культурной перспективе и культурная специфичность в обучении языку»).

as well as semantic and aspectual classes of predicates where it occurs. It also considers subtle semantic, pragmatic and communicative differences associated with non-negated and negated construction, respectively.
Such quasi-synonymy occurs primarily in cases when the predicate belongs to the aspectual class of accomplishments and denotes a telic process or action with a desired result ('to boil', 'to cool down', 'to warm up', 'to grow up', 'to finish', etc.). Those predicates include two major semantic components, that is, a lasting process or action and an instant result. In the imperfective aspect they allow at least two possible interpretations, namely, of a process and that of a result. Similar interpretations of sentences with such predicates occur due to different scope assignments of negation and *dolgo*. In sentences with non-negated predicate *dolgo* has scope over the 'process' component in the verb; in sentences with negated predicate negation has scope over the 'result' component of the verb while at the same time falling into the scope of *dolgo*. The former type of sentences describes long-lasting processes, whereas the latter type describes long-awaited results, which pragmatically amount to the same thing.

Key words: verbal aspect, negation, quasi-synonymy, semantics, pragmatics, corpus data, parallel corpora, aspectual classes

1. Введение

В работе рассматривается парадоксальная ситуация: русские фразы с длительной конструкцией без отрицания и с отрицанием с некоторыми предикатами квазисинонимичны: предложения *Чайник не закипал очень долго* и *Чайник закипал очень долго* могут быть употреблены применительно к одной и той же ситуации, а именно — вода в чайнике в течение очень долгого времени нагревается, но все не доходит до точки кипения. Сам по себе этот факт известен (см., напр., обсуждение на сайте https://gzom.ru/kak-nado/chajnik-dolgo-ne-zakipaet-byvaet-li-ne-pustym-mestom/; однако возникает вопрос, каков механизм подобной квазисинонимии. Сходный эффект возникает и с другими предикатами: *Компот остывал долго » Компот долго не остывал, Трубка долго раскуривалась » Трубка долго не раскуривалась*. В примерах типа *Компьютер долго не загружается* и *Компьютер долго загружается* эта квазисинонимия выражена еще более явно (что также не осталось незамеченным http://ortheos.livejournal.com/46166.html).[2]

Целью работы является определить, как в точности возникает квазисинонимия, какие семантические и прагматические механизмы здесь задействованы, а также какие семантические и аспектуальные классы предикатов ее позволяют.

2. Семантические механизмы квазисинонимии

Механизм квазисинонимии, по-видимому, заключается во взаимодействии отрицания и наречия *долго* с разными компонентами смысла предиката

[2] Данный эффект можно считать разновидностью эффекта «незначащего отрицания», рассмотренного в [Шмелев 2009; Shmelev 2016]

закипать. Глагол *закипать* относится к особой разновидности аспектуального класса accomplishments (а именно — несобственно-инхоативным предикатам[3]). Как и другие accomplishments, он предполагает две фазы — длительный процесс, ведущий к результату, и моментальный результат. Во фразах без отрицания в сферу действия *долго* попадает компонент 'процесс' и НЕСОВ имеет процессное значение; во фразах с отрицанием в сферу действия *долго* попадает компонент 'результат' под отрицанием и НЕСОВ имеет результативное значение, т. е. выступает в качестве тривиального коррелята к глаголу СОВ *закипеть*, появляющегося в контексте отрицания.

Иными словами, первая фраза имеет значение 'процесс, ведущий к результату, длился долго' (имплицируется — 'результат долго не наступал'); вторая — 'результат долго не наступал' (имплицируется — 'процесс, ведущий к результату, длился долго').

Заметим, кстати, что эффекту квазисинонимии подвержены не все несобственно-инхоативные предикаты. В частности, он не возникает, если результат нежелателен: фраза *Хлеб долго не черствел* значит, что хлеб долго не начинал черстветь (а не то, что хлеб, постепенно черствея, долго не становился совсем черствым), а фраза ?*Хлеб долго черствел* вообще прагматически неадекватна. Так же устроена пара *Капуста долго не гнила* vs. ?*Капуста долго гнила*. Первая фраза значит, что процесс гниения долго не начинался, а вторая прагматически неадекватна, если только ее не интерпретировать как эмфатическое высказывание о длительном нежелательном положении дел: *Капуста долго гнила на складе, прежде чем ее выбросили на помойку*.

По-видимому, это прагматическое требование имеет общий характер и связано с импликатурами, задаваемыми семантикой длительной конструкции в сочетании с отрицанием: высказывания о длительном ненаступлении некоторой ситуации более естественны, когда эта ситуация является желательной и долгожданной, чем наоборот; ср. несколько большую естественность фраз типа *У них долго не было детей* (и возникает импликатура, что дети были желанны); *Он долго не чинил сарай*; *Он долго не выздоравливал* по сравнению с фразами типа *Его долго не выгоняли с работы*; *Она долго не заболевала*; ?*Он долго не разбивал посуду*.

Сходный эффект возникает с некоторыми другими глаголами со значением вхождения в состояние или вхождения в процесс, напр. *заживать*. Ср. пример из «Национального корпуса русского языка» (НКРЯ):

(1) *Много уколов морских ежей. Они есть в Турции, на Кипре, на побережье Хорватии, да практически везде. Их уколы очень болезненны и долго заживают*. [Александр Мельников. Поцелуй медузы. Опасности

[3] Термин из статьи [Зализняк, Шмелев 2002]. В некоторых типологических работах используются несколько иные классификации; в частности, в работе [Татевосов 2010: 20] предикаты типа *закипать* относятся к акциональному типу «расширенных ингрессивно-непредельных» предикатов со значением 'вхождение в процесс'; предикаты типа *замерзать* классифицируются как «сильные стативно-процессные» со значением 'вхождение в состояние'.

подстерегают наших туристов даже в цивилизованных странах (2001) // «Известия», 2001.08.02]

Очевидно, что в последнюю фразу можно вставить *не* практически без изменения смысла (*Их уколы очень болезненны и долго не заживают*). Ср.:

(2) *При добыче кораллов плавать приходилось в кедах, тренировочном костюме и перчатках, потому что от коралловых царапин остаются долго не заживающие ранки.* [Александр Городницкий. «И жить еще надежде» (2001)]

Похожим образом ведет себя и глагол *сохнуть*[4] (ср. *Белье долго сохло* и *Белье долго не сохло*). Ср. примеры из НКРЯ:

(3) *Кожаные куртки для пешехода-туриста не годятся. Они тяжелы, громоздки, не пропускают воздуха к телу, а потому способствуют потению. Промокнув, долго сохнут.* [В. Семеновский. Снаряжение туриста (1929)]

(4) *Странные русского изготовления фотографические пластинки: после проявления и закрепления долго не сохнут!* [П. К. Козлов. Географический дневник Тибетской экспедиции 1923–1926 гг. №5 (1926)]

Необходимо при этом отметить, что коммуникативная и просодическая структура фраз с отрицанием и без него несколько различается: во фразах с отрицанием главное фразовое ударение падает на глагол, который является акцентоносителем ремы: *Чайник долго не ↓закипал.* Во фразах без отрицания главное фразовое ударение падает на наречие, и акцентоносителем ремы является именно оно: *Чайник ↓долго закипал*. Данное различие в коммуникативных структурах часто маркируется инверсией: *Чайник закипал ↓долго.*

При этом для описания подобных ситуаций, в зависимости от предиката, может чаще использоваться конструкция с отрицанием (семь примеров на *долго не закипать* и три примера на *долго закипать*, сорок восемь примеров на *долго не засыпать* и шесть примеров на *долго засыпать* в НКРЯ по состоянию на февраль 2018), конструкция без отрицания (единичные пример на *Компьютер долго не загружается* и тысячи примеров на *Компьютер долго загружается* в поисковике Google по состоянию на февраль 2018), или же обе конструкции могут быть представлены приблизительно в равной степени (31 пример в НКРЯ по состоянию на февраль 2018 на *долго не заживать* и 29 на *долго заживать*).

[4] Вообще, такая квазисинонимия свойственна многим глаголам, которые обозначают «изменение состояния, сопровождающееся становлением какого-то признака» [Гловинская 2001: 92–94]; ср. примеры из цитированной работы — *выздоравливать*, *замерзать*, *засыхать*, *оттаивать*, *согреваться* и др., — а также их подробный семантический анализ там же.

3. Сочетаемость конструкций *долго* и *долго не* с разными типами предикатов

Очевидно, что для того, чтобы точно установить все классы предикатов, для которых возможна описываемая квазисинонимия, необходимо определить область пересечения сочетаемости конструкций *долго* и *долго не*. При этом естественно ожидать, что большая часть сочетаемостных классов не совпадет — поскольку в некотором смысле *долго* и *долго не* антонимичны.

Анализ корпусных данных, в частности скетчей Sketch Engine (корпуса RuTenTen и Araneum Russicum Maius), сочетаемости RuSkELL, биграмм и триграмм в НКРЯ, дает следующее распределение основных сочетаемостных классов[5]:

***Долго* + глагол**

1. Непредельные действия и состояния (activities и states)

   Долго ждать <сидеть, смотреть, думать, работать>; долго мучиться <болеть, терпеть>

2. Предельные действия (accomplishments)

   Долго решать <уговаривать, подбирать, заводить, успокаивать, заканчивать>

3. Предельные процессы с желательным результатом

   Долго сохнуть <остывать, заживать, выздоравливать, загружаться, заводиться>

4. Непредельные процессы с желательным результатом

   Долго расти <худеть, набирать вес>

***Долго не* + глагол**

1. События и моментальные действия (achievements)

   Долго не приходить <не стрелять>

2. Предельные действия (accomplishments)

   Долго не обдумывать <не уговаривать>

3. Непредельные действия и состояния (activities и states)

   Долго не работать <не ждать, не сидеть, не размышлять, не думать>; долго не мучиться <не болеть, не терпеть>

---

[5] Здесь отражены не все возможные сочетания *долго* с разными типами предикатов, а только самые частотные; например, сочетание *долго гнил*, *долго кис*, *долго черствел* возможно в определенных прагматических условиях (см. выше), но не является типичным для *долго*.

4. Фазовые глаголы разных аспектуальных классов

Долго не начинать <не начинаться>; долго не кончать <не кончаться, не заканчивать, не заканчиваться>

5. Пердуративы разных аспектуальных классов

Долго не продержаться <не простоять, не проработать, не прослужить>

6. Модальные глаголы

Долго не мочь <не хотеть>

7. Предельные процессы с нежелательным результатом

Долго не вянуть <не черстветь, не засыхать, не изнашиваться>

8. Предельные процессы с желательным результатом

Долго не сохнуть <не заживать, не выздоравливать, не загружаться, не заводиться, не остывать>

9. Непредельные процессы с желательным результатом

Долго не расти <не худеть, не набирать вес>

Необходимо отметить, что конструкция *долго не* возможна во всех случаях, где возможна конструкция с *долго*, однако, естественно, далеко не везде ее употребление порождает квазисинонимию. Обратное неверно: с некоторыми предикатами (например, моментальными) сочетается конструкция с *долго не*, но в силу естественных семантических причин не сочетается конструкция с *долго* (**Он долго приходит*).

Начнем с тех случаев, когда такая квазисинонимия возможна. Она возникает в первую очередь у предельных процессов с желательным результатом (*сохнуть, остывать, заживать, выздоравливать, загружаться, заводиться*). Иногда предельность и желательность могут задаваться контекстом у непредельных глаголов и тогда для них становится возможен такой тип квазисинонимии: *нагреваться/охлаждаться до нужной температуры*, *набирать необходимый вес*. Почему она возможна?

Как было сказано выше, ее допускают чисто семантические механизмы сфер действия. Фразы типа *долго сохнуть* (где *долго* относится к компоненту ‘процесс’ в значении глагола) указывают на то, что процесс, который приводит к желательному результату, длится дольше, чем ожидалось. Фразы типа *долго не сохнуть* (где *долго* включает в свою сферу действия *не* и относится к компоненту ‘результат’ в значении глагола) указывают на то, что желательный результат долго не наступал. Видовременные свойства подобных предикатов также допускают эту квазисинонимию, т. к. она возникает отчасти за счет разной аспектуальной интерпретации глагола — *сохнуть* с *долго* употребляется в процессном значении НЕСОВ, а *сохнуть* с *долго не* — в результативном. Прагматические свойства данных предикатов также способствуют возникновению квазисинонимии. Дело в том, что сама по себе длительная конструкция с *долго*

предполагает, что нечто происходит дольше, чем ожидалось; этот, чисто количественный, вид оценки легко переходит в качественную, а именно — нечто происходит дольше, чем хотелось бы. Процесс, ведущий к желательному результату, происходит дольше, чем хотелось бы (*долго сохнуть*) — это прагматически то же самое, что желательный результат не наступает дольше, чем хотелось бы (*долго не сохнуть*).

Интересно, что некоторые глаголы из этих классов могут описывать и процессы с нежелательным результатом. При интерпретации нежелательности квазисинонимия пропадает. Во фразе (5) *долго* и *долго не* квазисинонимичны, но во фразе (6) *долго не* прагматически странно заменить на *долго*:

(5) *Компот очень долго остывал* ≅ *Компот очень долго не остывал*
(желательно, чтобы компот остыл),

но:

(6) *Этот чайник хорош тем, что вода в нем долго не остывает,*

при странности:

(7) ?*Этот чайник хорош тем, что вода в нем долго остывает.*

Таким образом, *долго не* может характеризовать длительное ненаступление нежелательного результата, что подтверждается и сочетаемостью с соответствующими предикатами (*долго не вянуть, долго не черстветь*), а *долго* редко характеризует длительное течение нежелательного процесса; ср. странность ?*Хлеб долго черствел* (о прагматических условиях, в которых допустимо подобное сочетание, уже говорилось выше).

Для непредельных процессов (типа *расти, худеть*) тоже возможно некоторое подобие квазисинонимии, но она имеет несколько другой механизм и носит более далекий характер, поэтому между фразами с *долго* и *долго не* сохраняются различия. *Долго* определяет течение процесса, а *долго не* — его начало, а не результат (результата нет, поскольку процессы непредельные). Ср. примеры с глаголом *расти*:

(8) *К сожалению, после стрижки у меня долго растут волосы*

(9) *После химиотерапии волосы долго не росли*

В обоих случаях речь идет о желательном росте волос, однако в примере (8) речь идет скорее о медленном росте уже имеющихся волос, а в примере (9) — скорее о том, что волосы долго не появлялись.

При этом для *долго не* возможна и интерпретация нежелательного роста:

(10) *После лазерной эпиляции волосы долго не растут*
(= 'нежелательные волосы долго не появляются').

Еще одна возможная сфера пересечения интерпретаций *долго* и *долго не* — фазовые глаголы типа *начинаться, кончаться* и пр. Квазисинонимия возможна в тех случаях, когда предикат может описывать не только точечный результат, но также длительный процесс или действие, что возможно далеко

не для всех предикатов и не для всех контекстов. В НКРЯ встретился единственный (весьма авторский) пример на *долго* в сочетании с фазовым глаголом, где сочетание *долго кончаться* квазисинонимично сочетанию *долго не кончаться*:

(11) *Все кончалось, кончалось и никак не могло кончиться, длилось, длилось и никак уже не могло продлиться, и кончалось бесконечно долго, кончалось мгновенно, длилось вечно, кончаясь всегда* [Александр Кабаков. Последний герой (1994–1995)].

В поисковике Google встретилось некоторое количество примеров, где квазисинонимия возникает в контекстах с глаголом *заканчивать*: *Леонардо долго заканчивал «Тайную вечерю» » Леонардо долго не заканчивал «Тайную вечерю», Один из плагинов долго заканчивал работу » Один из плагинов долго не заканчивал работу.* При этом многое в интерпретации определяется прагматикой: если во втором примере, где речь идет о нецеленаправленных агенсах (компьютерных программах), синонимия практически полная, то в первом может быть две разных ситуации. А именно, *долго не заканчивать* может описывать близкую к *долго заканчивать* ситуацию непрерывной работы над картиной, где очень растянулся конец, но может и относиться к другой ситуации — так сказать, «отложенного конца», когда художник прервал работу над картиной перед ее конечной стадией и долго не приступал к этой последней стадии. Во втором случае квазисинонимии с *долго заканчивать* не возникает.

Для некоторых предикатов наблюдается следующее семантическое распределение: фразы с *долго* без отрицания описывают свойства, а фразы с *долго не* — актуальные ситуации; ср.

(12) () *Этот кальян долго раскуривается.*

(13) () *Кальян долго не раскуривался.*

Интересно, что это отмечается носителями языка; ср. обсуждение на сайте https://rus.stackexchange.com/questions/6545/%D0%A7%D0%B0%D0%B9%D0%BD%D0%B8%D0%BA-%D0%BE%D1%81%D1%82%D1%8B%D0%B2%D0%B0%D0%B5%D1%82-%D0%B8%D0%BB%D0%B8-%D0%BD%D0%B5-%D0%BE%D1%81%D1%82%D1%8B%D0%B2%D0%B0%D0%B5%D1%82.

Есть класс предикатов, которые по своим аспектуальным свойствам подобны только что рассмотренным, однако не допускают квазисинонимии в силу семантических и прагматических причин. В частности, сюда относятся предикаты со значением природных процессов, не предполагающих человеческого участия, например *светать* и *темнеть*. Они свободно сочетаются с *долго не*, но не с *долго*, поскольку без отрицания данная конструкция предполагает какое-то участие человека в ситуации, хотя бы и не прямое; ср. некоторую странность ?*Летом долго темнеет*, ?*Зимой долго светает*[6].

Итак, почему квазисинонимия возможна именно у процессов? Структура события у предельных процессов типа *долго (не) сохнуть* примерно такова:

[6] Впрочем, подобные фразы встречаются в Интернете и воспринимаются как приемлемые некоторыми носителями языка.

'процесс ведет к некоторому результату', при этом в сферу действия *долго* попадает компонент 'процесс', а в сферу действия *долго не* — компонент 'результат'.

Для непредельных процессов в таком случае устанавливается прагматический предел: *долго худел*, *долго рос* — 'долго длился процесс, ведущий к достижению запланированного или желательного состояния'

Сочетание *долго не худел*, *долго не рос* маркирует отложенное н а ч а л о процесса, поэтому квазисинонимия носит несколько другой характер и воспринимается как более далекая.

У предикатов, относящихся к разновидностям states и activities, нет д в у х компонентов в значении, которые необходимы для установления квазисинонимии — процесса и результата.

Однако возникает вопрос: почему же не появляется квазисинонимия у предельных действий типа каузатива *успокаивать*, в структуре которых есть указание на длительное действие и его результат?

(14) *Она долго не успокаивала ребенка ≠ Она долго успокаивала ребенка*

Первая фраза указывает на то, что человек долго не приступал к действию (а не на то, что результат долго не наступал), а вторая — на то, что он долго его производил. Это происходит потому, что у целенаправленных и контролируемых каузативов маркировано начало действия, и именно к этому компоненту присоединяется *долго не*. Хотя у непредельных процессов *долго не* тоже определяет отложенное начало процесса, однако у процессов структура события в принципе более размытая из-за того, что они не инициируются человеком напрямую, и поэтому возможно возникновение более далекой квазисинонимии вида 'процесс долго не начинается' ≈ 'процесс долго идет'.

Заметим, что возможны прагматические условия, при которых некоторое подобие квазисинонимии может возникнуть у предикатов, которые, казалось бы, относятся к целенаправленных и контролируемых каузативам. Рассмотрим фразы *Он долго отвечает на письма* и *Он долго не отвечает на письма*. Если мы находимся на точке зрения субъекта, фразы отчетливо различаются по значению: первая указывает на то, что субъект, приступив к ответу на письма, тратит на это много времени, а вторая — что субъект долго не приступает к ответу на письма. Однако если мы встанем на точку зрения адресатов, которые ждут ответа, то различие может стать нерелевантным: адресат долго ждет и не может дождаться ответа. Адресатам неизвестно, забыл ли писатель писем про них вообще или просто долго пишет, и, с их точки зрения, данная ситуация может в равной степени описываться обеими фразами, которые тем самым становятся прагматически эквивалентными. Это же имеет место в некоторых неопределенно-личных конструкциях *Заказ очень долго доставляют* и *Заказ очень долго не доставляют*, с точки зрения заказчика, практически одно и то же: он в течение долгого времени не может получить свой заказ. Точно так же для матери, ожидающей, пока ее ребенок доест суп, фразы *Ребенок долго доедает суп* и *Ребенок долго не доедает суп* могут быть прагматически эквивалентны: ей безразлично, откладывается ли результат из-за того, что ребенок ест слишком медленно, или из-за того, что он отвлекся и вообще не приступает к доеданию.

Остальные группы предикатов не дают эффекта квазисинонимии. Хотя во всех случаях употребления *долго* это наречие может быть заменено на *долго не*, интерпретации, которые порождает вариант с отрицанием, совершенно другие. В них можно выделить два варианта — коммуникативно нейтральный и коммуникативно маркированный.

Первый характерен в первую очередь для целенаправленных действий типа *решать* или *работать*, независимо от их аспектуального класса (accomplishments или activities). Для сочетаний типа *долго не решал, долго не работал* более естественна интерпретация долго не начинающегося действия: долго не приступал к решению, долго не начинал работать. В такой интерпретации глагол под отрицанием входит в сферу действия наречия, причем *долго не* и глагол составляют единую рему, где акцентоносителем является глагол.

В сочетании с обозначениями состояний более естественна другая коммуникативная структура и другая сфера действия отрицания. Фразы типа *Он долго не мучился* более естественно интерпретировать как ‘Он мучился недолго’, где в сферу действия отрицания попадает контрастно рематизированное и акцентно выделенное наречие *долго*[7].

Это различие связано с разными коммуникативными ожиданиями: если естественно ожидать от человека работы, решения задач и пр., то ожидать, что он будет мучиться, менее естественно — это требует более расширенного контекстного обоснования. Соответственно, при *мучиться* вне специального контекста затруднено употребление отрицания; ср. навязанную контекстом интерпретацию с глаголом в сфере действия отрицания во фразах типа *Он выпил болеутоляющее и долго — часов пять — не мучился* (да и то приемлемость такой фразы признается не всеми носителями русского языка).

Среди примеров на *долго не* интерпретация ‘недолго’ характерна для целого класса предикатов, а именно, для пердуративов: *долго не прослужит* ‘прослужит недолго’, *долго не продержится* ‘продержится недолго’ и т. п. Именно на примере пердуративов можно наблюдать естественное (антонимичное) соотношение между интерпретациями *долго* и *недолго*: *долго продержится* указывает на то, что нечто будет иметь место в течение длительного времени (*долго* — рема), *долго не продержится* — на то, что нечто будет иметь место в течение недлительного времени (*долго* под отрицанием — рема). Фразы вида *Он долго работал* не антонимичны фразам вида *Он долго не работал*; последние обозначают длительное невыполнение действия, а не краткое действие. Это обусловлено тем, что у пердуративов имеется сильная семантическая валентность на временной отрезок. Слово *долго* заполняет эту валентность и, тем самым, оказывается тесно связано с предикатом и попадает в сферу действия *не*.

Как отмечалось выше, эффект квазисинонимии не возникает у агентивных целенаправленных глаголов типа *решать*: фраза *Он долго не решал задачу* означает, что субъект долго не приступал к ее решению, а это никак не синонимично

[7] Ср. обсуждение похожего явления на материале длительной конструкции в работе [Апресян 1995:68–70] на примерах вида *Крейсер не плавал два года*, где возможны две интерпретации — отрицательная (‘В течение двух лет крейсер не плавал’) и уменьшительная (‘Крейсер плавал меньше двух лет’).

фразе *Он долго решал задачу*, означающей, что субъект приступил к решению и долго не мог достичь результата (*долго не мог решить*). Фраза из «Капитанской дочки» Пушкина *…долго не распечатывал я пакета* означает, что Гринев в течение долгого времени колебался, распечатать ли пакет (ср. перевод *Long I hesitated to break the seal…* [Alexander Pushkin. *Marie: a Story of Russian Love* (Marie H. de Zielinska, 1877)]), тогда как фраза *долго распечатывал я пакет* означала бы, что субъект приступил к распечатыванию, но долго не мог справиться с задачей.

Мы видим, что у агентивных предикатов фразы с глаголом несовершенного вида с отрицанием и без отрицания не синонимичны; однако фраза с глаголом несовершенного вида без отрицания квазисинонимична фразе с глаголом совершенного вида и модальным глаголом *мочь* (*долго решал* означает, что *долго не мог решить*).

Снятие агентивности меняет интерпретацию фразы с отрицанием: *Задача долго не решалась* в естественном понимании значит, что попытки ее решить продолжались долго. В следующем примере из НКРЯ, в котором отрицание *не* стоит в скобках (наше видоизменение исходного текста), без обращения к оригиналу нельзя точно установить, была ли исходная фраза с *не* или без *не*:

(15) *Дело это задерживалось из-за того, что долго (не) решался вопрос о новой форме обмундирования — будет ли оно темно-зеленого или защитного цвета?* [А. Ф. Редигер. История моей жизни (1918)]

Так же обстоит дело со множеством аналогичных предикатов, напр. *успокаивать-успокаиваться*: декаузативы с отрицанием и без отрицания квазисинонимичны. Ср. примеры из НКРЯ:

(16) *Я слез, разревелся еще раз и медленно, долго успокаивался.* [Василий Белов. Плотницкие рассказы (1968)]

(17) *Там он не мог сдержаться и плакал при виде жены ли, матери, другой родни, чувствуя себя во всем виноватым. И долго не успокаивался.* [Борис Екимов. Высшая мера (1995)]

Таким образом, здесь возникает имеем дело с другой парадоксальной квазисинонимией, а именно, активы и каузативы без отрицания имеют приблизительно ту же интерпретацию, что и пассивы и декаузативы с отрицанием: *Мать долго успокаивала ребенка* предполагает, что *Ребенок долго не успокаивался*.

Почему же снятие агентивности приводит к возникновению квазисинонимии? Потому что в декаузативе маркирован результат, а не начало, и *долго не* присоединяется к этому компоненту значения.

В большинстве случаев декаузативы (как и рассмотренные выше инхоативы) при описании соответствующих ситуаций употребляются с отрицанием. Так, в НКРЯ (по состоянию на февраль 2018) находим пять примеров употребления глагола *заводиться* в сочетании с наречием *долго* (в значении, как в предложении *Машина долго не заводилась*) с отрицанием и ни одного без отрицания.

При этом естественно возникает импликатура, что в конечном счете результат был достигнут. В этом отношении показательны данные переводных текстов, напр. (из НКРЯ):

(18) *Я чувствовал себя почти рассерженным и долго не мог успокоиться. — I felt almost vexed, and was long in calming myself.* [Ivan Turgenev. Annouchka (Franklin P. Abbott, 1884)]

(19) *…спички еще долго не отыскивались. — …it was a long time before the matches were found.* [Fedor Dostoevsky. The Possessed, or The Devils (Constance Garnett, 1913)]

(20) *Нежданов долго не соглашался… — It took Nejdanov a long time before he consented…* [Ivan Turgenev. Virgin Soil (Rochelle S. Townsend, 1929)]

(21) *Но он долго не мог уснуть… — But it was long before he could sleep.* [Leo Tolstoy. The Awakening (parts 2–3) (William E. Smith, 1900)]

Не случайно в похожих контекстах переводчики на русский язык часто вставляют в перевод *не*, хотя отрицание отсутствовало в оригинале (особенно часто — при наличии модального глагола), напр.:

(22) *It took me quite a while to get to sleep… — Я долго не засыпал…* [Дж. Д. Сэлинджер. Над пропастью во ржи (Р. Райт-Ковалёва, 1965)]

(23) *Took me a while to figure out what I wanted to do. — Долго не могла понять, что мне нравится.* [Майкл Коннели. Город костей (Д. Вознякевич, 2006)]

Особняком в ряду рассмотренных явлений стоит пара *долго решался — долго не решался*. Это тоже квазисинонимия, однако она имеет другую причину, связанную с тем, что *решаться* и *не решаться* обозначают в некотором смысле одно и то же, а именно — пребывать в состоянии нерешительности, которое завершается или не завершается тем, что человек *решился*[8]. Характерно использование оборота *долго не решаться* в переводных текстах, напр. (из НКРЯ):

(24) *She hesitated a long time. — Она долго не решалась начать.* [Джон Стейнбек. Гроздья гнева (Н. Волжина, 1940)]

Таким образом, рассмотренный в настоящей статье эффект обусловлен взаимодействием ряда факторов: аспектуальным классом предиката, его семантикой и, в очень большой степени, прагматическими особенностями ситуации и знаниями о мире.

[8] Подробнее см. [Булыгина, Шмелев 1997: 173; Шмелев 2015: 202–203].

Литература

1. *Апресян Ю. Д.* (1995). Избранные труды: В 2 т. Т. II: Интегральное описание языка и системная лексикография. М.: Школа «Языки русской культуры».
2. *Булыгина Т. В., Шмелев А. Д.* (1997). Идентификация событий. Онтология, аспектология, лексикография // Булыгина Т. В., Шмелев А. Д. Языковая концептуализация мира (на материале русской грамматики). М., 1997. С. 167–179.
3. *Гловинская М. Я.* (2001). Многозначность и синонимия в видо-временной системе русского глагола. М.: Азбуковник.
4. *Зализняк Анна А., Шмелев А. Д.* (2002). Семантика начала с аспектологической точки зрения // Логический анализ языка. Семантика начала и конца. М., 2002. С. 212–224.
5. *Татевосов С. Г.* (2010). Акциональность в лексике и грамматике: автореферат диссертации… доктора филолог. наук. Московский государственный университет им. М. В. Ломоносова, Москва 2010.
6. *Шмелев А. Д.* (2015). Имперфективация и видовая корреляция // Зализняк Анна А., Микаэлян И. Л., Шмелев А. Д. Русская аспектология: в защиту видовой пары. М.: Языки славянской культуры, 2015. С. 192–207.
7. *Шмелев А. Д. «Незначащее» и «невыраженное» отрицание* (когнитивные и коммуникативные источники энантиосемии) // Логический анализ языка. Ассерция и негация. М., 2009. С. 173–202.
8. *Shmelev A.* (2016), Semantic shifts as sources of enantiosemy, The lexical typology of semantic shifts, Cognitive Linguistic Research, vol. 58, Mouton de Gruyter, Berlin—Boston, p. 67–93.

References

1. *Apresjan Ju. D.* (1995), Selected works, in 2 vol., vol. 2: Integrated description of language and systematic lexicography [Izbrannye trudy, v 2 t., t. 2: Integral'noe opisanie yazyka i sistemnaya leksikografiya], Shkola «Yazyki Russkoi Kul'tury», Moscow.
2. *Bulygina T. V., Shmelev A. D.* (1997) Identification of events: ontology, aspectology, lexicography [Identifikatsiya sobytii. Ontologiya, aspektologiya, leksikografiya], Linguistic conceptualization of the world (based on the Russian grammar) [Yazykovaya kontseptualizatsiya mira (na materiale russkoi grammatiki)], Shkola «Yazyki Russkoi Kul'tury», Moscow.
3. *Glovinskaya M. Ya.* (2011) Polysemy and synonymy in aspect-tense system of Russian verb [Mnogoznachnost› i sinonimiya v vido-vremennoi sisteme russkogo glagola], Azbukovnik,.
4. *Shmelev A. D.* (2009), "Void" and "unexpressed" negation (cognitive and communicative sources of enantiosemy) ["Neznachashchee" i "nevyrazhennoe" otritsanie (kognitivnye i kommunikativnye istochniki enantiosemii)], Logical analysis of language: assertion and negation [Logicheskii analiz yazyka: assertsiya i negatsiya], Moscow, p. 173–202.
5. *Shmelev A. D.* (2015), Imperfectivization and aspectual correlation [Imperfektivatsiya i vidovaya korrelyatsiya], Zaliznyak Anna A., Mikaelyan I. L., Shmelev A. D. Russian aspectology: in defense of aspectual pair [Russkaya aspektologiya: v zashchitu vidovoi pary], Yazyki slavyanskoi kul›tury, Moscow, p. 192–207.
6. *Shmelev A.* (2016), Semantic shifts as sources of enantiosemy, The lexical typology of semantic shifts, Cognitive Linguistic Research, vol. 58, Mouton de Gruyter, Berlin—Boston, p. 67–93.
7. *Tatevosov S. G.* (2010), Actionality in lexicon and grammar (doctoral thesis, manuscript) [Aktsional›nost› v leksike i grammatike: avtoreferat dissertatsii… doktora filolof the start og. nauk], Moscow State University, Moscow.
8. *Zaliznyak Anna A., Shmelev A. D.* (2002), Semantics of the start viewed through aspectology [Semantika nachala s aspektologicheskoi tochki zreniya], Logical analysis of Language: Semantics of the start and end [Logicheskii analiz yazyka: semantika nachala i kontsa], Moscow, p. 212–224.

РАЗРЕШЕНИЕ НЕОДНОЗНАЧНОСТИ СФЕР ДЕЙСТВИЯ В ПИСЬМЕННЫХ ТЕКСТАХ (НА МАТЕРИАЛЕ АНГЛИЙСКОГО ЯЗЫКА)[1]

Апресян В. Ю. (valentina.apresjan@gmail.com, vapresyan@hse.ru)

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

DISAMBIGUATION OF SCOPE IN WRITTEN ENGLISH TEXTS

Apresyan V. Ju. (valentina.apresjan@gmail.com, vapresyan@hse.ru)

National Research University "Higher School of Economics", Moscow, Russia

The paper is a corpus study of the factors involved in disambiguating potential scope ambiguity in written sentences with negation and universal quantifier *all*, such as *I cannot visit all these universities*, which, depending on topic-focus assignment, can alternatively mean 'I cannot visit any of these universities' (*cannot* is focus) and 'I cannot visit some of these universities' (*all* is focus). The factors at play in scope disambiguation are the syntactic function of the constituent containing *all* (subject, direct complement, adjunct); the status of the main predicate and *all* with respect to the information structure of the utterance (topic vs. focus); veridical vs. nonveridical context; sentence type (unreal conditional, rhetorical question); and pragmatic implicatures pertaining to the situations described in the utterances. The paper also demonstrates differences in the frequency distribution of various scope readings and their underlying causes, as well as formulating typical contexts for each scope interpretation.

Key words: scope, ambiguity, disambiguation, negation, universal quantifier, information structure, topic, focus, veridicality, implicature

[1] Публикация подготовлена в ходе проведения исследования по проекту «Factors in resolving scope ambiguity» (№ 18-01-0007) в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2018–2019 гг. и в рамках государственной поддержки ведущих университетов Российской Федерации «5–100».

1. Введение

Разрешению лексической и синтаксической неоднозначности посвящено огромное количество работ, в то время как разрешение неоднозначности сфер действия (СД) семантически активных операторов начало привлекать к себе внимание исследователей сравнительно недавно (ср. [Kadmon, Roberts 1986] о просодии и СД, [Kurtzman, MacDonald 1993] о разрешении неоднозначных СД, [Koizumi, 2009] о прагматике и просодии при анализе неоднозначных СД). При этом сам феномен такой неоднозначности изучен весьма подробно, особенно в формальной лингвистике; ср. лишь некоторые из работ в этой области [Hintikka 1973, Ioup 1975, Cooper 1979, Aoun, Li 1989, Horn 1989, Богуславский 1996, Reinhart 1997, Kiss 2006 и другие]. Данная работа имеет целью заполнить эту лакуну.

Интересно, что в случае потенциальной неоднозначности СД, как и в случаях с лексической и синтаксической неоднозначностью, сами носители языка редко испытывают затруднения с определением правильной интерпретации. Приведем пример. Так, английское наречие *accidentally* ‘нечаянно, случайно’ может иметь разные сферы действия над группой глагола — либо широкую, над всей группой, как в примере (1), либо узкую — только над зависимым глагола, как в примере (2) (примеры взяты из корпуса EnTenTen на ресурсе Sketch Engine):

(1) *He accidentally* [*burned his thumb*] *while cooking*
Он случайно [обжег себе большой палец]

(2) *I accidentally planted* [*beets*] *in the garden one year*
Однажды я случайно посадил в огороде [свеклу]

В примере (1) случайным является как ожог, так и его локализация (так называемый purely accidental agent), в то время как в примере (2) само действие *посадить* намеренно и контролируемо, а случайным является лишь посадка свеклы вместо какого-то другого овоща — например, потому, что сажавший перепутал семена (так называемый mistaken intentional agent[2]). Определение сферы действия *accidentally* в каждом из случаев не представляет затруднений и диктуется прагматическими соображениями. Действие *обжечься* маловероятно в качестве неслучайного, т. е. намеренного и контролируемого, в то время как действие *посадить* невозможно в качестве ненамеренного и неконтролируемого; поэтому в первом случае случайной является вся возникшая ситуация, а во втором — лишь один из ее аспектов. Таким образом, можно сформулировать правило определения СД *accidentally* с разными типами предикатов:

(3) Если глагол указывает на намеренное контролируемое действие, *accidentally* имеет узкую СД — только над глагольным аргументом: [*Innocent civilians*] *are being accidentally bombed* ‘Бомбы были случайно сброшены на [мирных жителей]’ [Сама бомбардировка намеренна, попадание в мирных жителей случайно]

[2] См. [Martin 2015] о различиях между purely accidental и intentional mistaken agents.

(4) Если глагол указывает на ненамеренное неконтролируемое действие, *accidentally* имеет широкую СД — над всей глагольной группой:
He accidentally [*spilled tea on her skirt*] ‘Он случайно/нечаянно [пролил чай ей на юбку]’ [Само проливание было случайным]

Если глагол указывает на частично контролируемое действие, т. е. на такое, где результат не контролируем, а предшествующее ему действие может быть как намеренным, так и случайным, то возможны обе СД, и для установления правильной интерпретации требуются экстралингвистические факторы:

(5) *He accidentally* [*shot Mary*]
‘Он случайно/ нечаянно [выстрелил в Мэри]’
(ружье случайно выстрелило)

(6) *He accidentally shot* [*Mary*]
‘Он случайно выстрелил в [Мэри]’
(случайно попал в Мэри, а целился в Джона)

Интересно, что это различие, которое в английском передается при помощи разных СД, в русском языке лексикализовано. Наречие *нечаянно* описывает только полностью незапланированные и неконтролируемые действия: *нечаянно толкнуть/уронить/разбить/задеть* и пр., но не ?*нечаянно посадить свеклу вместо морковки,* ?*нечаянно обстрелять мирных жителей*. В то же время, *случайно* может описывать как полностью ненамеренные и неконтролируемые действия (*случайно толкнуть/уронить/разбить/задеть*), так и намеренные действия с неконтролируемым результатом: *случайно обстрелять мирных жителей* (человек намеренно стрелял), *случайно оказаться в темном переулке* (человек намеренно передвигался), *случайно обнаружить записку в ящике стола* (человек намеренно открыл ящик стола). Безусловно, между *нечаянно* и *случайно* есть и другие различия — например, обязательное указание на нежелательный характер действия и в то же время на его не слишком значительный масштаб у *нечаянно*.

2. Цели работы и материал исследования

Целью настоящей работы является установление всех факторов, в том числе и прагматических, способствующих разрешению неоднозначности СД отрицания. В качестве объекта исследования выбраны фразы с отрицанием и квантором всеобщности в английском языке, поскольку они представляют собой один из наиболее частых и хорошо известных случаев неоднозначности сфер действия. Ср. следующий пример, допускающий два разных прочтения в зависимости от СД отрицания: *I did not read all the bad news*.

(7) а. *I did not read* [*all*] *the bad news*[3]
‘букв. Я не прочел [все] плохие новости’
‘Я прочел не все плохие новости’

б. *I did not* [*read*] *all the bad news*
‘букв. Я не [читал] все плохие новости’
‘Я не читал всех этих плохих новостей’

Отметим, что в русском переводе различию в сферах действия могут соответствовать аспектологические и падежные различия, передающие различия в референции (когда *all* входит в СД отрицания, референция конкретная и определенная, когда не входит — конкретная, но неопределенная).

Во втором типе интерпретации, где *all* не входит в СД отрицания, возможен еще один вариант прочтения. А именно, в СД отрицания может входить не матричный глагол, а какая-либо другая составляющая, при ее наличии в контексте; ср.

в. *I did not read all this bad news* [*merely to get upset*]
‘букв. Я не читал все плохие новости
[только для того, чтобы расстроиться]’
‘Я читал все эти плохие новости не для того, чтобы расстроиться’
(а с другой целью)

В данном случае отрицание является контрастным и рематизирован отрицаемый целевой компонент.

Исследование проведено на материале письменных текстов, поскольку при чтении отсутствуют некоторые «подсказки», которые имеются при устном восприятии, и носитель языка должен опираться на дополнительные индикаторы. Так, при устном произнесении правильную интерпретацию часто позволяет установить просодия, которая маркирует коммуникативное членение

---

[3] Существует традиция описывать СД отрицания в примерах, где отрицание воздействует на количественную группу, как широкую, т. е. включающую в себя глагол и его зависимое; ср. анализ примера в [Апресян 1995:68–70]: *Крейсер не* [*плавал два года*] = ‘Крейсер плавал меньше двух лет’ (≈ ‘Крейсер плавал не два года’). Однако представляется, что в английских примерах с *all* типа (7а) более естественно и информативно описывать СД отрицания как включающую не всю группу глагола, а только рематизированный квантор всеобщности: *I did not read* [*all*] *the bad news*, давая интерпретацию ≈ ‘I read some of the bad news’. При этом при широкой СД отрицания и при СД отрицания, включающей только *all*, интерпретации прагматически равнозначны: ‘Неверно, что я читал все плохие новости’ ≈ ‘Я читал не все плохие новости’, хотя чисто логически интерпретация с широкой СД возможна и в ситуации, когда человек не читал ничего. Из этих соображений в примере (7а) и далее СД отрицания с рематизированным квантором *all* описывается как узкая и включающая только *all*, но не глагол.

высказывания, являющееся одним из главных факторов в интерпретации сфер действия[4].

В работе использованы данные параллельного англо-русского и русско-английского корпуса НКРЯ. Выбор параллельного подкорпуса НКРЯ в качестве отправной точки для исследования был неслучайным, поскольку его использование позволяет верифицировать правильность интерпретации: во многих случаях в русских оригиналах или в русских переводах с английского интерпретация однозначна. Непосредственным объектом исследования послужили оригинальные и переводные английские тексты, в то время как русский материал используется для целей верификации правильности интерпретации СД отрицания.

Поисковый запрос формулировался следующим образом: *not* + Verb + *all*, с расстоянием до 3 слов. Все нерелевантные контексты, в частности идиоматические выражения с фиксированной интерпретацией (*not at all*, *all of it, at all costs, after all, all the more*) были вручную исключены из результатов поиска[5].

В итоге для формулировки изначальных гипотез было рассмотрено 147 контекстов из НКРЯ. Исследование позволило внести существенную ясность в вопрос разрешения СД отрицания, в том числе, установить частотное распределение интерпретаций, факторы, влияющие на него, а также круг индикативных контекстов, способствующих разрешению неоднозначности в письменных текстах. Естественное в данном контексте экспериментальное исследование использования говорящими контекстов-индикаторов для разрешения неоднозначности СД отрицания осталось за рамками данной работы, однако планируется в будущем.

3. Результаты исследования по данным параллельного корпуса НКРЯ

Распределение контекстов в НКРЯ оказалось следующим:

- 82 контекста, где отрицание *not* имеет СД над квантором *all*, например *You haven't told me* [*all*] » 'Ты не сказал мне всего » Ты не все мне сказал';

---

[4] См. [Sgall, Hajičová, Benesová 1973], [Partee 1991] и [Hajičova 1998] об актуальном членении предложения и сфере действия отрицания, [Ionin 2001] об актуальном членении предложения и СД кванторов, а также [Kadmon, Roberts 1986] о роли просодии в определении неоднозначных СД.

[5] Поскольку исследование носит пилотный характер, мы не рассматривали контексты с отрицательными местоимениями, которые также допускают разные СД с квантором всеобщности; ср. *Nobody can be there* [*all*] *the time* 'Никто не сможет быть там в течение ВСЕГО времени' [СД местоимения над квантором] vs. *Nobody* [*will be there*] *all this time* 'В течение всего этого времени, там никого не будет' [СД местоимения над группой глагола]. Безусловно, в будущем планируется рассмотреть также контексты с отрицательными местоимениями.

- 58 контекста, где отрицание *not* имеет СД над матричным глаголом, например *I don't* [*care*] *about all these idiots* » 'Я не интересуюсь всеми этими идиотами » Я не интересуюсь никакими из этих идиотов', или над другой составляющей, *I didn't come all the way from Alabama* [*to hear you say that*] » 'Я проделала весь этот путь из Алабамы, не для того, чтобы такое от тебя услышать';
- 7 неоднозначных контекстов.

Некоторые контексты, неоднозначные в английском, могут быть интерпретированы при помощи русского оригинала. Ср. следующий пример из перевода *Анны Карениной* на английский:

(8) It all seemed to her [Darya Alexandrovna] a far simpler matter: all that was needed, as Marya Philimonovna had explained, was to give Brindle and White breast food and drink, and *not to let the cook carry all the kitchen slops* to the laundry maid's cow (Leo Tolstoy, *Anna Karenina*, trans. Constance Garnett, 1911).

Английское предложение может быть интепретировано как тотальный запрет при СД *not* над глаголом (нельзя забирать *никакой* части помоев), либо как частичный при СД *not* над *all* (нельзя забирать *все* помои). Однако соответствующее русское предложение, в котором отсутствует квантор всеобщности, естественно интерпретировать как тотальный запрет: «Ей (Дарье Александровне) казалось все это гораздо проще: что надо только, как объясняла Матрена Филимоновна, давать Пеструхе и Белопахой больше корму и пойла и чтобы повар *не уносил помои из кухни* для прачкиной коровы» (Л. Н. Толстой, «Анна Каренина»).

Таким образом, неоднозначность сфер действия, требующая широкого контекста для разрешения, — достаточно редкое явление. Кроме того, частотное распределение интерпретаций предполагает, что в английском языке случаи, когда квантор *all* не входит в СД отрицания, являются более редкими и, соответственно, более маркированными семантически, прагматически и коммуникативно.

3.1. Факторы, влияющие на интерпретацию СД отрицания

СД отрицания определяется следующими взаимосвязанными, но не взаимозаменимыми факторами:

- Коммуникативная структура высказывания (входит ли *all* в тему или в рему);
- Семантика (входит ли *all* в пресуппозицию);
- Синтаксис (наличие «конкурирующей» составляющей, которая может перетягивать на себя СД отрицания);
- Иллокутивный тип высказывания (утверждение, восклицание, риторический вопрос, условие);
- Прагматика (наличие импликатур в высказывании).

Кратко проиллюстрируем взаимосвязь разных факторов. Ср. фразы *I haven't read all these stupid books!* 'Я не читал всех этих глупых книг' [СД отрицания над глаголом] и *I haven't read all the books* 'Я прочитал не все книги' [СД отрицания над квантором]. Различия между этими фразами касаются

1. Коммуникативных структур: в первой рема глагол, во второй — квантор;
2. Семантических структур: в первой квантор входит в пресуппозицию, во второй — в ассерцию;
3. Иллокутивного типа высказывания: первая фраза содержит эмфазу, вторая — нет, первая — восклицание, вторая — утверждение.

Вдобавок к перечисленному, есть некоторые дополнительные синтаксические соображения. Так, если *all* входит в подлежащее, то часто предпочтительным способом отрицания является отрицание при составляющей, а не при глаголе. Так, фразы типа (9б) существенно частотнее фраз типа (92):

(9) а. ?*All Russians are not gloomy* [в интерпретации с СД отрицания над *all*]
б. *Not all Russians are gloomy* (Vladimir Nabokov, *Pale Fire*)

Однако при контрастном отрицании на кванторе подобные фразы становятся возможными. Они задают определенный тип коммуникативной структуры (*all*-составляющая в контрастной теме) и СД отрицания (над квантором *all*); в корпусе встречаются отдельные фразы такого типа:

> [*All*] *men are not confirmed old bachelors like me and the Colonel* 'Не все мужчины такие убежденные старые холостяки, как мы с полковником' [Bernard Shaw. Pygmalion (1912)].

Однако в целом фразы такого рода настолько редки в параллельном корпусе НКРЯ, что в данной работе отдельно не рассматривались.

Чаще всего *all* попадает в СД отрицания, будучи частью прямого дополнения, как в примере (10):

(10) *He didn't like all his students* ['Ему нравились не все его студенты']

Если квантор *all* является частью обстоятельства, ему проще «ускользнуть» от отрицания:

(11) *He didn't talk to me all these years* ['Он не говорил со мной все эти годы']

Поскольку решающим фактором для определения СД отрицания является актуальное членение, в дальнейшем обсуждении для классификации разных возможных случаев используется именно оно — оно практически однозначным образом задает СД отрицания. При этом показывается, каким образом различные синтаксические, семантические и прагматические факторы влияют на выбор той или иной коммуникативной структуры (и, таким образом, на интерпретацию СД отрицания), а также каким образом изменение этих параметров меняет актуальное членение высказывания.

3.2. Коммуникативная структура высказывания

Внизу рассматриваются три типа структур — квантор и глагол в теме, квантор в теме, глагол в реме и глагол в теме, квантор в реме. На рассмотренном материале не встретилась четвертая, логически возможная структура: и глагол, и квантор в реме. Такая структура возможна, например, для прохибитивов; ср.

Don't spend all the money 'Не трать все деньги', где отрицание имеет СД над квантором. Представляется, что для прохибитивов маловероятны интерпретации, где отрицание имеет СД над глаголом, вида *Don't touch all these sensitive issues* 'Не затрагивай все эти щекотливые темы'; скорее, в такого рода контекстах будет использовано отрицательно поляризованное местоимение *any*: *Don't touch any of these sensitive issues* 'Не касайся никаких из этих щекотливых тем'.

3.2.1. Квантор *all* и глагол в теме

Когда *all* и матричный глагол входят в тему высказывания, они обычно являются частью пресуппозиции и, соответственно, не попадают в СД отрицания (см. обсуждение в [Sgall, Hajičová, Benesová 1973, Hajičová 1998]). Таким образом, в СД отрицания попадает какая-то другая составляющая, обычно приглагольная — дополнение или обстоятельство:

(12) *I am not paying all your bills* ‖ [TO MAKE YOU THINK YOU CAN SPEND MORE][6]
'букв. Я не плачу твои счета [чтобы ты думал, что можно тратить больше]' »
'Я плачу твои счета не для того, чтобы ты думал, что можно тратить больше'

В примере (12) факт оплаты счетов упоминается в теме, а предполагаемая (и отрицаемая) причина — в реме. Однако данное актуальное членение устанавливается на прагматической основе: интерпретация вида 'Я плачу не все твои счета, чтобы ты тратил больше' не допускает рационального осмысления. При замене последующего контекста меняется коммуникативная структура и, соответственно, СД отрицания:

(13) *I am not paying* ‖ [ALL] ‖ *your bills because I have no money*
'Я не плачу все твои счета, поскольку у меня нет денег'

Пример (13) может быть интерпретирован и как тотальный отказ от уплаты счетов (*all* в теме и не входит в СД отрицания), и как согласие платить лишь часть (*all* в реме и входит в СД отрицания). Контексты типа (13), где и квантор, и глагол в теме, а в СД отрицания входит какой-либо другой рематический элемент, являются верификативными (Янко 2011) и носят полемический и контрастный характер. В силу этого они весьма редко встречаются. В нашей выборке из 147 контекстов подобная структура встретилась лишь дважды (1.5%); ср. пример:

(14) Robert, you know as well as I do that the Priory has not protected the truth all these years [Тема] || [TO HAVE IT GATHER DUST UNTIL ETERNITY] [Рема] (Dan Brown, *The Da Vinci Code* (2003)) » '… Аббатство защищало правду все эти годы [Тема] || не [для того, чтобы она вечно покрывалась пылью] [Рема]'

3.2.2. *All* в теме, глагол в реме

В подобных предложениях квантор *all* входит в тему, а матричный глагол в рему. Обычно глагол при этом либо не фактивный, либо помещен

[6] Здесь и далее квадратные скобки в примерах обозначают СД отрицания, а знак || — границу между темой и ремой.

в невериdикативный контекст (будущее время, вопрос и пр.; см. [Giannakidou 1998, Падучева 2011]); ср. пример, где глагол употреблен в значении будущего:

(15) I'm not [GOING] [Рема] all the way to Huntingdon to celebrate the ruby wedding of two people I have spoken to once for eight seconds since I was three [Тема] (Helen Fielding, *Bridget Jones's Diary*, 1996] » 'Я не [собираюсь] [Рема] проделывать весь путь в Хантингтон, чтобы отпраздновать сорокалетие свадьбы людей, который я видел секунд восемь, когда мне было три года [Тема]'.

В силу того, что глагол в подобных высказываниях не фактивен, он не входит в пресуппозицию и, соответственно, попадает в СД отрицания. Интересно, что замена контекста на веридикативный дает сдвиг СД отрицания с глагола (который перемещается в тему) на целевое обстоятельство:

(16) I have not come all the way to Huntingdon [Тема] || [TO CELEBRATE THE RUBY WEDDING OF TWO PEOPLE I HAVE SPOKEN TO ONCE FOR EIGHT SECONDS SINCE I WAS THREE] [Рема] » 'Я проделал весь этот путь в Хантингтон [Тема] не [для того чтобы отпраздновать сорокалетие свадьбы людей, который я видел секунд восемь, когда мне было три года] [Рема] (а с какой-то другой целью)'.

Предложения типа (15), где глагол входит в рему, встречаются существенно чаще, чем предложения типа (16), где и глагол, и *all* входят в тему: они составляют 38 % от всех рассмотренных контекстов (56 из 147). Возможно, это связано с тем, что они не столь контрастивны, как контексты типа (16). Часто индикатором контекстов типа (15) является наличие показателя определенности при кванторе *all*, особенно указательного местоимения *this*, *these*, что способствует топикализации квантора и выведению его из СД отрицания. Кроме того, выведению квантора из СД отрицания способствует эмфаза (*all the stupid, all the blooming, all the goddamn* и т. п.). Ср. примеры (17) и (18), где введение эмфатического компонента при *all* меняет коммуникативную структуру фразы и СД отрицания:

(17) *I didn't understand* [*all the questions*] *he asked me* ['Я понял не все его вопросы'].

(18) *I didn't* [*understand*] *all these stupid questions he asked me* ['Я не понял всех этих тупых вопросов']

3.2.3. *All* в реме, глагол в теме

В данном типе предложений глагол и вся прочая часть высказывания входят в тему и не входят в СД отрицания, в то время как составляющая, содержащая *all*, формирует контрастную рему и входит в СД отрицания; ср.:

(19) The right rim of the casket had not fallen [Тема] || [ALL] [contrastive Focus] the way to the floor [Рема] and was still propped partially on its supports '(Dan Brown, *Angels and Demons* (2000)) 'Правый край гроба упал [Тема] || не [до конца] [рема] и все еще отчасти опирался на подставки'

Данный тип коммуникативного членения и ассоциированная с ним СД отрицания встречается наиболее часто (82 контекста из 147 или 56 %). Известное

высказывание, приписываемое Аврааму Линкольну, принадлежит именно к данному типу:

(20) *You can fool all the people some of the time, and some of the people all the time, but you cannot fool* [*all the people all the time*] (Abraham Lincoln) ‘Можно всё время дурачить некоторых, можно некоторое время дурачить всех, но нельзя всё время дурачить всех’

4. Коммуникативный тип высказывания

Некоторые типы высказываний ассоциированы с определенным типом интерпретации СД отрицания. Так, восклицательные предложения с указательными местоимениями, где *all* используется для эмфазы, а не в качестве квантора, способствуют выводу квантора из СД отрицания, и помещению в него глагола: *I can’t stand* [*listening*] *to all these pompous speeches!* » ‘Не могу слушать все эти помпезные речи’. Подобной интерпретации также способствует контекст ирреального условия и риторического вопроса:

(21) My […] visit to London—would have been altogether delightful, [HAD] I not [been OVERWHELMED] all the time by anxiety, impatience, anguished forebodings (Vladimir Nabokov. *Look at the Harlequins!* (1974)) ‘Мой визит в Лондон был бы чудесным, если бы я не [был] все время [подавлен беспокойством, нетерпением, плохими предчувствиями’

(22) “How many acres?” “About fifteen.” “Why not [sow] all?” cried Levin. (Leo Tolstoy. Anna Karenina (Parts 1–4). (Trans. Constance Garnett (1911)) ‘«Сколько акров?» «Примерно пятнадцать». «Почему бы не [засеять] все?» крикнул Левин’

Интересно, что в примере (22) соответствующее прочтение с *all* вне СД отрицания возникает именно в переводе; в оригинале квантор *весь*, без всякого сомнения, входит в СД отрицания, поскольку *не* стоит не при глаголе, а при составляющей, содержащей квантор:

— А клевер? — Послал Василия с Мишкой, рассевают. Не знаю только, пролезут ли: топко. — На сколько десятин? — На шесть. — Отчего же не [*все?*] *— вскрикнул Левин* Л. Н. Толстой. «Анна Каренина»).

Если добавить в эту фразу глагол, отрицание становится риторическим, как и в английском переводе, и имеет СД над всей пропозицией: *Отчего же не* [*засеять все*] » ‘Нужно засеять все; говорящий спрашивает, почему адресат этого не хочет’.

Существует естественное объяснение тому факту, что ирреальные условия и риторические вопросы являются неблагоприятными контекстами для отрицания квантора. Первые содержат пресуппозицию (или, согласно [Karttunen and Peters 1975], импликатуру) того, что имеющая место ситуация противоположна той, что вводится условием: фраза вида *Если бы он пришел…*

подразумевает, что он не пришел. Соответственно, когда высказывание с ирреальным условием содержит отрицание, оно имплицирует, что пропозиция без отрицания истинна. Соответственно, находящийся в ней квантор всеобщности сам по себе недоступен для отрицания, которое вместо этого присоединяется ко всей пропозиции (*Если бы ты не* [*поссорился со всеми друзьями*]*, сейчас тебе было бы с кем поговорить*).

Риторические вопросы с отрицанием содержат подобную импликатуру; хотя они не предполагают того, что противоположная пропозиция непременно истинна, в той же степени, что ирреальные условные предложения, они тем не менее передают уверенность говорящего в том, что имеет место ситуация, противоположная той, о которой он спрашивает. Ср. риторический вопрос *Не любил ли я тебя все эти годы?*, подразумевающий убежденность говорящего в истинности ситуации, которая находится под отрицанием. Соответственно, отрицанию также подвергается вся пропозиция.

Ирреальные условия и риторические вопросы составляют 20 % процентов от всех предложений, где в СД отрицания находится глагол (11 из 56)[7].

5. Прагматические импликатуры

В некоторых контекстах, когда все прочие факторы позволяют разные чтения, интерпретация СД отрицания осуществляется на основе прагматических факторов. Ср. следующую пару примеров:

(23) *I haven't* [SLEPT] *all night* ['Всю ночь я не спал']

(24) *I haven't slept* [ALL] *day* ['Я спал не весь день (а лишь часть)']

Относительно дня и ночи в культуре существуют разные прагматические ожидания: люди обычно спят большую часть ночи, и совсем не спят днем. Сообщение о том, что кто-то спал часть ночи, а часть ночи не спал (при СД отрицания над квантором всеобщности) было бы прагматически неинформативным в смысле кооперативного принципа Грайса, т. к. такая ситуация является обычной (люди спят большую часть ночи). Соответственно, во фразах, где речь идет о сне в течении ночи для отрицания выбирается СД над глаголом на основе прагматических соображений. Фразы вида *Я не спал всю ночь* встретились в рассмотренной выборке 18 раз, исключительно с интерпретацией СД отрицания над глаголом.

С другой стороны, фразы вида (24) получают прагматически естественную интерпретацию ('Я спал часть дня, но не весь день'), если в СД отрицания входит квантор. Это полемические фразы с контрастной ремой, предполагающие, что в предтексте кто-то утверждал, что субъект спал *весь* день.

---

[7] На русском материале описано большое количество контекстов плеонастического отрицания, не ведущего себя как отрицательный оператор; ср., например, контексты вида *разве не, неужели не*, а также отрицание в вопросах-просьбах [Булыгина, Шмелев 1997], контексты с глаголами ожидания вида *Жду, пока он не придет* [Барентсен 1980]. Во всех подобных контекстах отрицание имеет СД над всей пропозицией и не может относиться к квантору всеобщности.

В общем случае, интерпретация СД отрицания с темпоральными выражениями вида *X не Y весь Z*, где Z — обозначение отрезка времени, — зависит от семантического типа предиката, и регулируется следующими прагматическими правилами. Для моментальных глаголов действует следующее правило:

(25) Необычно, если действие, обозначаемое глаголом Y, ни разу не происходит в течение временного отрезка Z

Соответственно, под отрицание попадает глагол: *He had not* [THOUGHT] *of her all evening* 'Он не [думал о ней] весь вечер' » 'За вечер он ни разу о ней не подумал'; *They* [HAD] *not* [SPOKEN] *all day* 'Они не [говорили] весь день' » 'За весь день они ни разу не заговорили друг с другом'.

Для дуративов действует следующее правило:

(26) Необычно, если действие, обозначаемое глаголом Y, происходит в течение всего временного отрезка Z

Соответственно, под отрицание попадает квантор всеобщности: *The concert couldn't have detained you* [ALL] *this time* 'Ты не мог быть на работе [все это время]', *I simply cannot work* [ALL] *the time* 'Я не могу работать [все время]'.

6. Заключение

Таким образом, можно видеть, что разрешение потенциально неоднозначных сфер действия — это многоуровневый механизм, в котором взаимодействуют разные лингвистические факторы (синтаксические, семантические, прагматические), а временами требуется и экстралингвистическая информация. Количественные результаты исследования суммированы в таблице 1 в Приложении.

Из таблицы видно, что маркированным вариантом интерпретации в английском языке является *not V* [*all*], т. е. предпочтительна сфера действия отрицания над квантором, а не над глаголом. Такая интерпретация не только встречается более редко (58 примеров на СД отрицания над глаголом vs. 82 примера на СД отрицания над квантором), но и возникает в особых условиях. Как видно из таблицы, лишь в семи примерах такого рода из 58 отсутствует специальный контекст и интерпретация определяется только знанием ситуации, без помощи семантических, синтаксических, сочетаемостных, а также конвенционализованных прагматических факторов. Во всех прочих примерах представлен один из контекстов-индикаторов — а) пунктивный предикат в контексте временного обстоятельства (*He didn't ring all weekend*), б) контекст риторического вопроса или ирреального условия (*Had I not waited for you all this time?*, *As if I hadn't waited for you all this time*), в) контекст *sleep all night* (*He didn't sleep all night*), г) контекст эмфатического указательного местоимения *all this, all that* (*I don't want to listen to all this*).

Напротив, установление СД отрицания над квантором практически не ассоциировано ни с какими специальными индикативными контекстами и определяется естественными прагматическими ожиданиями и знанием ситуации. Единственный вполне, хотя и не стопроцентно, индикативный контекст для

интерпретации *not* [*V*] *all* — предикаты со значением длительной ситуации при вхождении *all* в состав обстоятельства времени (*I hadn't spend all the winter in Africa*). В остальных примерах количественная (а не эмфатическая) интерпретация *all* и, следовательно СД отрицания над квантором, устанавливаются на контекстуальных прагматических основаниях. Есть некоторые сочетаемостные индикаторы того, что группу квантора следует интерпретировать как неполное количество — а именно, контекст предикатов возможности и долженствования (*I couldn't read all the books, You needn't read all the books*), контекст ментальных предикатов, предикатов физического восприятия и речи (*I didn't understand/know all, I didn't see/hear/say all*). Однако ни один из них не исключает возможности интерпретации СД отрицания над глаголом.

Литература

1. *Апресян Ю. Д.* (1995). Избранные труды: В 2 т. Т. II: Интегральное описание языка и системная лексикография. М.: Школа «Языки русской культуры». [Apresjan Ju. D. Izbrannye trudy: V 2 t. T. II: Integral'noe opisanie yazyka i sistemnaya leksikografiya [Selected works: In 2 vol. Vol. II: Integrated description of language and systematic lexicography]. Moscow: Shkola «Yazyki Russkoi Kul'tury», 1995.]
2. *Барентсен А.* (1980). Об особенностях употребления союза пока при глаголах ожидания. Studies in Slavic and General Linguistics 1, 17–68. Amsterdam: Rodopi. [Barentsen A. J., Ob osobennostjax upotreblenija sojuza poka pri glagolax ozhidanija 'On the peculiarities of using the conjunction poka with the verbs of expectation. Studies in Slavic and General Linguistics 1, 17–68. Amsterdam: Rodopi.]
3. *Богуславский И. М.* (1996). Сфера действия лексических единиц. [Boguslavsky, I. M. Sfera dejstvija leksicheskix edinic 'Scope of lexical items'. Moscow: Shkola «Yazyki Russkoi Kul'tury», 1996.]
4. *Булыгина Т. В., Шмелев А. Д.* (1997). Языковая концептуализация мира (на материале русской грамматики). М.: Школа «Языки русской культуры». [Bulygina T. V., Shmelev A. D. Yazykovaya kontseptualizatsiya mira (na materiale russkoy grammatiki 'Linguistic conceptualization of the world (on the material of the Russian grammar)'. Moscow. Shkola "Jazyki russkoy kul'tury", 1997].
5. *Падучева Е. В.* (2011). Имплицитное отрицание и местоимения с отрицательной поляризацией. Вопросы языкознания, 1. 2011. С. 3–18. [Paducheva E. V. Implicitnoe otricanie i mestoimenija s otricatel'noj poljarizaciej 'Implicit negation and negative polarity pronouns'. Vorposy jazykoznanija 'Problems of linguistics', 1. P. 3–18.
6. *Aoun, J., and Y. A. Li* (1989). Scope and Constituency. Linguistic Inquiry 20–2: 141–172.
7. *Bartsch R.* (1973) «Negative Transportation» gibt es nicht // Linguistische Berichte. Bd. 23.
8. *Giannakidou, A.* (1998) Polarity Sensitivity as (Non)veridical Dependency. John Benjamins, Amsterdam-Philadelphia.

9. *Hajičova E.* (1998) Topic-Focus Articulation, Tripartite Structures and Semantic Content, Kluwer, Dordrecht.
10. *Hintikka, J.* (1973). Quantifiers vs. quantification theory. Dialectica, 27:329–358. Reprinted in Linguistic Inquiry 5 (1974):153–177.
11. *Horn, L. R.* (1989). A Natural History of Negation. University of Chicago Press, Chicago.
12. *Kadmon, N., Roberts, C.* (1986). Prosody and scope: The role of discourse structure. CLS Proceedings.
13. *Ionin T.* The one girl who was kissed by every boy: Scope, scrambling and discourse function in Russian. // Proceedings of ConSole X. 6580.
14. *Ioup, G. L.* (1975). The Treatment of Quantifier Scope in a Transformational Grammar. Diss. City U. of New York. Jackendoff, R. 1972. Semantic Interpretation in Generative Grammar. Cambridge, Mass.: MIT Press.
15. *Karttunen L., Peters S.* (1975) Conventional implicature in Montague grammar. In BLS 1: Proceedings of the First Annual Meeting of the Berkeley Linguistics Society. Berkeley, California. Pp. 266–278.
16. *Koizumi, Y.* (2009). Processing the not-because ambiguity in English: the role of pragmatics and prosody. CUNY thesis.
17. *Kiss, K. É.* (2006). Quantifier Scope Ambiguities. In: The Blackwell Companion to Syntax. Everaert, M. and H. van Riemsdijk (eds).
18. *Kurtzman, H. S., and MacDonald, M. C.* (1993). Resolution of quantifier scope ambiguities. Cognition, 48, 243–279.
19. *Martin, F.* (2015) Explaining the link between agentivity and non-culminating causation. In Proceedings of Semantics and on non-culminating interpretations of telic predicates 75 Linguistic Theory (SALT) 25, p. 246–266.
20. *Partee, Barbara H.* (1991). Topic, focus and quantification. In SALT I: Proceedings of the First Annual Conference on Semantics and Linguistic Theory 1991, eds. Steven Moore and Adam Zachary Wyner, 159–187. Ithaca, N.Y.: CLC Publications, Department of Linguistics, Cornell University.
21. *Reinhart, T.* (1997). Quantifier Scope: How Labour is Divided between QR and Choice Functions, Linguistics and Philosophy 20, 335–397.
22. *Sgall, P., Hajičová, E. & Benesová, E.* (1973). Topic, Focus and Generative Semantics. Kronberg, Taunus: Scriptor.

Приложение

Таблица 1. Список контекстов и интерпретаций сферы действия отрицания

Контекст	**not V [all] — в сфере действия отрицания квантор** **82 примера**	**not [V] all — в сфере действия отрицания глагол** **58 примеров**
с названиями временных периодов типа *all the time, all morning, all winter*	7 примеров, с предикатами, обозначающими длительные ситуации, типа *work* ‘работать’, *spend* ‘проводить’, *stay* ‘жить’	17 примеров, из них - 12 примеров с пунктивными предикатами, типа *ring* ‘позвонить’, *make the bed* ‘убрать постель’ - 5 примеров с предикатами, обозначающими длительные ситуации, в контекстах риторического вопроса и ирреальных условных клауз
в контексте *sleep all night*	—	18 примеров
в риторических вопросах и ирреальных условных клаузах (без обозначения временного периода)	—	6 примеров
в неколичественных эмфатических контекстах, включая контексты с *all this, all that*	—	17 примеров, из них 8 с *all this,* 2 с *all that*
в количественных не-эмфатических контекстах	75 примеров, из них - 18 примеров с *could* - 18 примеров с *know, understand, realize, guess* и пр. - 7 примеров с *say, tell*	—

HOW MUCH DOES A WORD WEIGHT? WEIGHTING WORD EMBEDDINGS FOR WORD SENSE INDUCTION

Arefyev N. (narefyev@cs.msu.ru)

Lomonosov Moscow State University & Samsung Moscow Research Center, Moscow, Russia

Ermolaev P. (permolaev@cs.msu.ru)

Lomonosov Moscow State University, Moscow, Russia

Panchenko A. (panchenko@informatik.uni-hamburg.de)

University of Hamburg, Hamburg, Germany

The paper describes our participation in the first shared task on word sense induction and disambiguation for the Russian language RUSSE'2018 [Panchenko et al., 2018]. For each of several dozens of ambiguous words, the participants were asked to group text fragments containing it according to the senses of this word, which were not provided beforehand, therefore the „induction" part of the task. For instance, a word "*bank*" and a set of text fragments (also known as "contexts") in which this word occurs, e.g. "*bank is a financial institution that accepts deposits*" and "*river bank is a slope beside a body of water*" were given. A participant was asked to cluster such contexts in the unknown in advance number of clusters corresponding to, in this case, the "company" and the "area" senses of the word "*bank*". The organizers proposed three evaluation datasets of varying complexity and text genres based respectively on texts of Wikipedia, Web pages, and a dictionary of the Russian language.

We present two experiments: a positive and a negative one, based respectively on clustering of contexts represented as a weighted average of word embeddings and on machine translation using two state-of-the-art production neural machine translation systems. Our team showed the second best result on two datasets and the third best result on the remaining one dataset among 18 participating teams. We managed to substantially outperform competitive state-of-the-art baselines from the previous years based on sense embeddings.

Keywords: lexical semantics, word sense induction, word sense disambiguation, neural machine translation, clustering, word embeddings

СКОЛЬКО ВЕСИТ СЛОВО? ВЗВЕШИВАНИЕ ВЕКТОРОВ СЛОВ В ЗАДАЧЕ ИЗВЛЕЧЕНИЯ ЗНАЧЕНИЙ СЛОВ

Арефьев Н. (narefyev@cs.msu.ru)
Московский Государственный Университет им. М. В. Ломоносова и Московский Исследовательский Центр Самсунг, Москва, Россия

Ермолаев П. (permolaev@cs.msu.ru)
Московский Государственный Университет им. М. В. Ломоносова, Москва, Россия

Панченко А. (panchenko@informatik.uni-hamburg.de)
Университет Гамбурга, Гамбург, Германия

Ключевые слова: лексическая семантика, извлечение смыслов слов, разрешение лексической многозначности, нейросетевой машинный перевод, кластеризация, векторные представления слов

1. Introduction

Word sense induction (WSI) task aims at identification of word senses for ambiguous words in unsupervised and knowledge-free manner i.e. without using any manually compiled dictionaries or sense inventories. While a few languages, such as English, have such lexical resources of good quality and coverage the appeal of the WSI setting is the possibility to enable word sense disambiguation (WSD) for languages and domains where such resources are not available. Slavic languages still do not have lexical resources with broad coverage comparable, for instance, to English WordNet [Miller et al, 1990] which provides a comprehensive inventory of senses. The word sense induction task was thoroughly studied in the context of a few popular Western European languages, such as English, French, and German. However, for the Russian language only few word sense disambiguation experiments were performed, e.g. [Lopukhina et al, 2016] motivating the need for more research in this field.

Main research results related to word sense induction and disambiguation were effectively reported on the materials of the English language. Notably, several shared tasks performed a systematic evaluation of approaches in this field. Namely, [Agirre and Soroa, 2007] presented a SemEval task where participants were provided with contexts in English which they had to group according to word senses. The gold standard annotation used WordNet sense inventory. [Manandhar et al., 2010] presented a similar evaluation campaign, which was devoted to word sense induction of nouns and verbs. For each target word, participants were provided with a training set in order to learn the senses of that word. Then, participating systems disambiguate

unseen instances (contexts) of the same words using the learned senses. [Jurgens and Klapaftis, 2013] performed an evaluation in the multi-sense labeling task. In this setup, participating systems provide a context with one or more sense labels weighted by the degree of applicability. More recently, [Alagić et al., 2018] presented an instance representation based on lexical substitutes—contextually suitable meaning-preserving replacements of words in context. [Corrêa et al., 2018] proposed a method that leverages recent findings in word embeddings research to generate context embeddings. Their word sense induction method represents a set of ambiguous words as a complex network, where edges are generated based on word embeddings similarity. [Pelevina et al., 2016] investigate another graph-based approach to word sense induction which relies on a graph clustering method applied to an ego-network of distributionally related words, which is constructed using word embeddings. [Panchenko et al., 2017] rely on a similar approach, making the induced senses interpretable using hypernymy labels, images, and definitions of senses extracted in an unsupervised way.

Some research related to word sense induction and disambiguation was also performed before for the Russian language, however not as a part of an evaluation campaign, but rather as individual contributions with often incomparable evaluation benchmarks, making it difficult to compare performance of different approaches. Loukachevitch and [Chuiko, 2007] proposed a method for all-word disambiguation task on the basis of a thesaurus. [Kobritsov et al., 2005] developed disambiguation filters to provide semantic annotation for the Russian National Corpus. The semantic annotation was based on the taxonomy of lexical and semantic facets. Lyashevskaya and [Mitrofanova, 2009] proposed a statistical word sense disambiguation models on an example of several nouns. [Lopukhin et al., 2017] evaluated several approaches: Adaptive Skip-gram, Latent Dirichlet Allocation, clustering of contexts, and clustering of synonyms. [Ustalov et al., 2017] proposed a meta-clustering algorithm for graphs designed for unsupervised acquisition of word senses and grouping them into synsets using Wiktionary and other dictionaries of synonyms.

In this paper we make a further step in this direction, improving upon the current state-of-the-art results. Our experiment is performed in a competitive setting of the first shared task on word sense induction and disambiguation RUSSE'2018 [Panchenko et al., 2018] aiming at comparing sense induction and disambiguation systems for the Russian language. Namely, we present two approaches to WSI for the Russian language. One of the approaches was the second best in two datasets and the third best in the remaining one dataset in this evaluation campaign where 18 participants submitted over a hundred of various models. Besides, our model was better that one of the state-of-the-art approaches to WSI based on AdaGram sense embeddings [Bartunov et al., 2016]; [Lopukhina et al., 2016].

The paper is structured the following way. First, in **Section 2**, we describe the positive result: an approach based on clustering of contexts represented as a weighted average of word embeddings. We used word2vec embeddings and compared different weighting schemes. This method yielded the best results. Second, in **Section 3**, we present a negative result: an alternative approach based on neural machine translation. Machine translation has improved a lot recently after the introduction of neural network based translation systems. In order to translate ambiguous words

correctly a translation system should disambiguate them first, so we hoped to benefit from translating contexts of ambiguous words from Russian into English with two of the best available machine translation systems, namely Yandex Translate and Google Translate. This approach showed worse results than the first one and we analyze possible reasons. Next, in Section 4, we present a qualitative analysis of both approaches on several examples. Finally, we conclude the paper with a summary of the experiments and contributions.

2. Word Sense Induction via Clustering of Contexts Represented as a Weighted Average of Embeddings

This section presents **a positive result**, using an approach to word sense induction, which managed to achieve highly competitive results in the shared task.

2.1. Description of the Method

In this approach, the word sense induction task is formalized as a clustering task. Namely, each context of an ambiguous word is represented as a weighted average of all of its words' embeddings with carefully selected weights. Finally, all contexts of the ambiguous word represented as vectors in the same space are clustered. Different clusters are interpreted as different senses of the ambiguous word. The three steps of the method are described below.

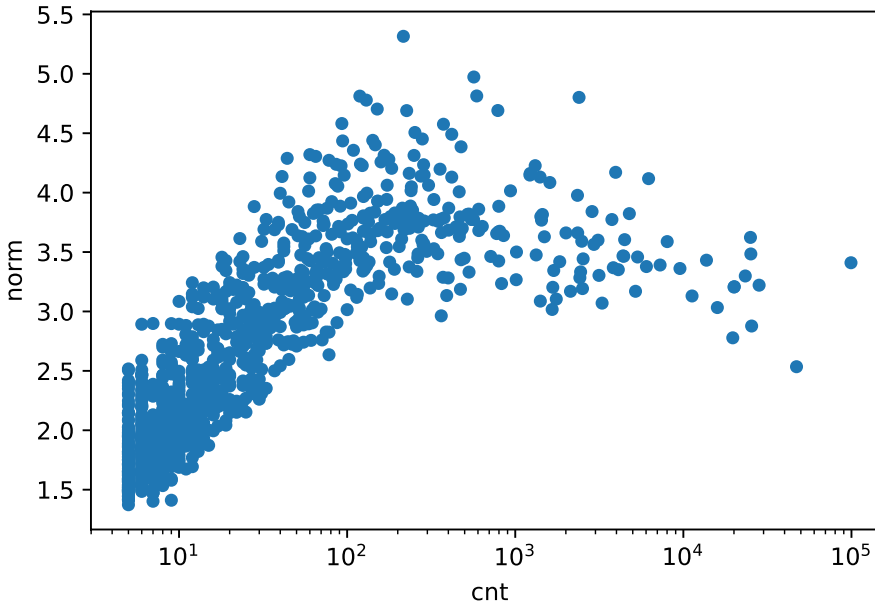


Figure 1. L2 norm of skip-gram word vectors as a function of word frequency: the higher the word frequency the larger the norm of this word embedding

2.1.1. Learning Word Embeddings from a Large Unlabeled Corpus

The train and the test datasets of the shared task are fairly small making the representations learned only from them perform poorly. For this reason, the information from the background collections is exploited in the form of word embeddings. Namely, we trained several word2vec models [Mikolov et al., 2013], both CBOW and Skip-gram, using different corpora: the corpus of books in Russian with 150 Gb of plain text extracted from lib.rus.ec library in [Arefyev et.al., 2015], [Panchenko et al. 2016] and the Russian Wikipedia containing about 3 Gb of plain text. Also, we experimented with various hyperparameters including window size, minimum word frequency, and corpora preprocessing type. During preliminary experiments on the training data, for active-dict and bts-rnc datasets, we have chosen the Skip-gram model with window size 10 and word embeddings of size 200 trained for 3 epochs on Librusec and containing only words with at least 5 occurrences (resulting in 3 million words vocabulary). For wiki-wiki dataset, we have selected a model with similar hyperparameters trained on the Russian Wikipedia which performed better for this dataset.

It is common to normalize word2vec embeddings to unit length when they are used for word similarity estimation. However, according to our experiments on the train datasets, unnormalized embeddings provided better results and we decided to perform no normalization. To explain this we plot the dependency between word frequency and corresponding embedding length for a random sample of 1 thousand words (cf. **Figure 1**). One can see a strong positive correlation with shorter embeddings for rare words and longer embedding for more frequent ones. This dependency does not hold only for the most frequent words which are few. Shorter embeddings affect weighted average less compensating for lower quality of embeddings for rare words which can explain better performance of unnormalized words vectors.

2.1.2. Representing Contexts in a Vector Space

To represent each example as a vector we calculated a weighted average of word-2vec embeddings for all context words, i.e. all words excluding the target (polysemous) word and all of its occurrences in all grammatical forms (this helped reducing noise introduced by different senses of the target word mixed in its embedding). An appropriate selection of the weights turned out to be critical for the overall performance as we show later. We experimented with unweighted average, tf-idf weights and chi-square statistic values (for shortness we use the term “chi2” from now on). Tf-idf weights were estimated on Russian Wikipedia. Tf-idf weights help lowering the importance of the most frequent words which are likely non-informative and should not affect the context vector much. However they do not help to differentiate between less frequent context words which are related to the target and the ones which appeared with the target by chance. To give the former more weight we used chi2 measuring independence between context words and target words. If some context word occurs frequently but only with a single target word it is natural to consider it a good feature for discriminating between this target word’s senses. Chi2 weights are higher for those context words which appear mostly with a particular target word compared to those context words which appear with different target words uniformly. For instance, high chi2 weights were given to context words like “*страхования*” (*insurance*) appearing mostly with the target word “*полис*” (*policy / city-state*), “*леса*” (*forest*) appearing

mostly with "*опушка*" (*woodside / trimming*) and so on. Our best results were achieved by raising to the properly selected powers and then multiplying tf-idf and chi2 weights. We experimented with normalization of the weight vector and the resulting weighted average vector and found that normalizing both of them using L2 norm works best.

2.1.3. Clustering of Word Context Vectors

Finally, for each target word separately we clustered all of its examples' context vectors. In the preliminary experiments, we tried different clustering algorithms including DBSCAN and its extensions HDBSCAN and OPTICS, Affinity Propagation, Spectral clustering, Agglomerative clustering as implemented in scikit-learn[1], pyclustering[2] and hdbscan[3] software libraries. We used only Affinity Propagation [Frey & Dueck, 2007] and Agglomerative clustering algorithms in the final experiments as they have shown the most promising results on the train datasets.

2.2. Experiments and Discussion of the Results

2.2.1. Optimization of Hyperparameters on the Train Sets

In our experiments, we tried different word embedding models, clustering algorithms and word weighting methods for the context words. For each word in a context we look up this word's embedding and multiply it by a weight. This weight is a product of tf-idf and chi2 weights raised to different powers which were selected individually for each dataset on the corresponding train set from 6 values between 0 and 2.5 (totally 36 combinations of powers).

Despite the variable number of senses per word in the datasets, Agglomerative clustering with a fixed number of clusters showed the best results for active-dict and bts-rnc. However, for wiki-wiki, the best performance was achieved using Affinity Propagation which determines the number of clusters automatically resulting in a different number of clusters for different words. The following hyperparameters and their values were evaluated on the train set of each dataset individually and the best performing values were used for the test set.

- **Agglomerative clustering** is a bottom-up hierarchical clustering approach, where each data point is placed in its own cluster first, and the the most similar clusters are merged as one moves up the hierarchy[4]. The method has three meta-parameters:
  - *number of clusters* (from 1 to 14, 2 by default)
  - *distance between points* (euclidean, L1, L2, manhattan, cosine; all of our best result use euclidean distance so we do not mention it further)
  - *linkage* criterion defines the distance between clusters (average, ward, complete, ward by default)

[1] http://scikit-learn.org

[2] https://github.com/annoviko/pyclustering

[3] https://github.com/scikit-learn-contrib/hdbscan

[4] https://en.wikipedia.org/wiki/Hierarchical_clustering

- **Affinity Propagation clustering** is based on the message passing metaphor and finds "exemplar" members representative of clusters. The algorithm chooses the number of clusters based on the provided input data. However, the algorithm has two meta-parameters[5]:
  - *damping* controls the probability for a point to change it's cluster (between 0.5 and 1, 0.5 by default);
  - *preference* affects the probability of creating a new cluster and hence the number of clusters (chosen between −20 and 5).

Table 1. Hyperparameters selected on the train sets when word2vec model trained on lib.rus.ec is used. The models in bold were ranked 2nd best in the final ranking (cf. **Table 3**)

Dataset	Clustering Algorithm	Clustering Hyperparameters	Word Weights	Train ARI
wiki-wiki	Agglomerative Clustering	n_clusters=2 linkage=ward	$tfidf^{1.5} * chi2$	0.8057
	Affinity Propagation	damping=0.5 preference=−6.8	$tfidf^{1.5} * chi2^{0.5}$	0.8148
bts-rnc	**Agglomerative Clustering**	**n_clusters=10 linkage=average**	$\mathbf{tfidf^{1.5} * chi2^{0.5}}$	0.2633
	Affinity Propagation	damping=0.9 preference=−2.9	$tfidf^{2} * chi2^{2}$	0.1448
active-dict	**Agglomerative Clustering**	**n_clusters=3 linkage=ward**	$\mathbf{tfidf^{1.5} * chi2^{0.5}}$	0.2535
	Affinity Propagation	damping=0.5 preference=−1.0	tfidf * chi2	0.2414

The results of hyperparameter selection for each of the best clustering algorithms when the Skip-gram word2vec model trained on lib.rus.ec is used are presented in **Table 1**. The second best result in the final ranking, according to the private test ARI score, was obtained on two of the datasets using these hyperparameters. However, on the wiki-wiki dataset simple average of word2vec vectors trained on lib.rus.ec without weights performed better according to the public test score so it was used for the final submission (see **Table 3** for the final ranking). It is interesting that the best results in **Table 1** were achieved with the same weighting scheme for all datasets (powers 1.5 and 0.5 for tf-idf and chi2 respectively) but with different clustering algorithms. Namely, Affinity Propagation was better than Agglomerative Clustering on wiki-wiki but worse on the other two datasets. This fact and also much better ARI scores on wiki-wiki can be explained by more coarse-grained senses in this dataset and hence better separability of context vectors which is necessary for algorithms like Affinity Propagation to select the correct number of clusters automatically.

[5] https://en.wikipedia.org/wiki/Affinity_propagation

Next, we investigate how much a properly selected weighting scheme affects the results. Performance of several variations of our method with different word weighting schemes is represented in Table 2. For each clustering algorithm and a weighting scheme the best weight powers and clustering algorithm hyperparameters were selected on the train sets. Tf-idf weights always outperform chi2 weights and their combination is always the best weighting scheme.

Figure 2 shows how ARI depends on the number of clusters and linkage for agglomerative clustering. Ward linkage with 2–3 clusters shows uniformly good results on all datasets. However for bts-rnc average linkage with 10 clusters performs little better.

Table 2. Impact of the weighting scheme and the clustering algorithm on performance

Method	Word Weights	wiki-wiki	bts-rnc	active-dict
Agglomerative Clustering	the best power score (cf. Table 1)	0.8057	0.2633	0.2535
	only tf-idf	0.7882	0.2618	0.2451
	only chi-squared score	0.7776	0.2355	0.2157
	no weights	0.6565	0.2237	0.2147
Affinity Propagation	the best power score (cf. Table 1)	0.8148	0.1448	0.2414
	only tf-idf	0.7623	0.1406	0.2335
	only chi-squared score	0.7525	0.1371	0.1908
	no weights	0.7866	0.1108	0.1950

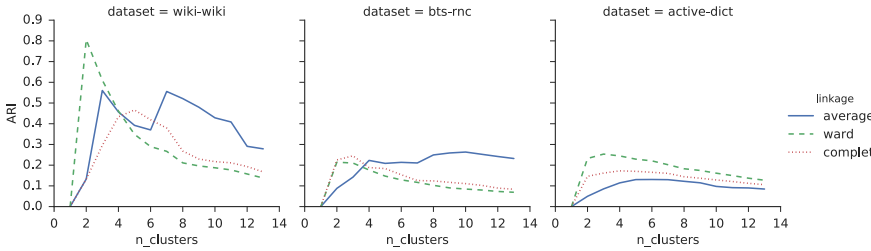


Figure 2. Performance of word sense induction on or the train sets depending on the number of clusters and the linkage for agglomerative clustering

Figure 3 shows a correlation between weights powers and ARI scores. For each dataset we plot a heatmap for the clustering algorithm which achieved the best score on that dataset with carefully selected hyperparameters and with the default ones. The figure shows that simple multiplication of tf-idf and chi2 weights does not work significantly better than just using tf-idf weights. It is important to properly adjust tf-idf and chi2 weights bringing them into the same scale to obtain really good improvement.

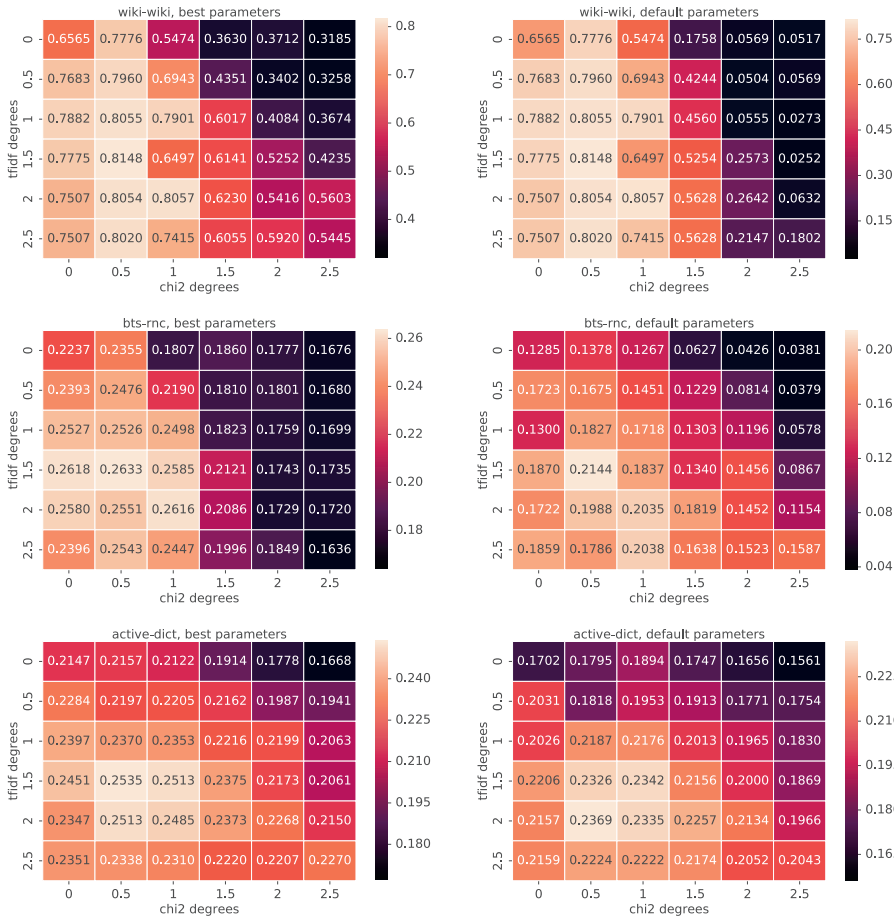


Figure 3. ARI on the train sets depending on powers of tf-idf and chi2 weights for the best clustering algorithm (cf. **Table 1**) with default hyperparameters and selected hyperparameters

2.2.2. Submitted Results for the Best Models Identified on the Train Set

Finally, Table 3 shows our best-submitted results. As one can observe, we obtain very competitive results scoring second on two datasets (bts-rnc and active-dict) and third on the wiki-wiki dataset. Furthermore, for the wiki-wiki dataset, where we ranked third, a difference with the second best participant is relatively small, as compared to the difference between the first and the second participants. We conclude that the developed methods are highly competitive with the state-of-the-art methods for word sense induction for the Russian language.

Table 3. The best results of our experiments, which were submitted to the RUSSE shared task. The final rank among 18 other participants is indicated in the round brackets

Dataset	Word2vec	Weights	Clustering	Train ARI	Test ARI (public)	Test ARI (private)
wiki-wiki	Wikipedia	—	Affinity Propagation	0.7577	1.0000	0.6586 (3)
	The second best submission in the shared task (akutuzov)			—	0.9823	0.7096 (2)
	The best submission in the shared task (jamsic)			—	1.0000	0.9625 (1)
bts-rnc	lib.rus.ec	$tfidf^{1.5}$*$chi2^{0.5}$	Agglomerative	0.2633	0.2812	0.2818 (2)
	The best submission in the shared task (jamsic)			—	0.3508	0.3384 (1)
active-dict	lib.rus.ec	$tfidf^{1.5}$*$chi2^{0.5}$	Agglomerative	0.2535	0.2361	0.2270 (2)
	The best submission in the shared task (jamsic)			—	0.2643	0.2477 (1)

3. Using Neural Machine Translation for Word Sense Induction and Disambiguation

This section presents **a negative result**. We exploited two state-of-the-art neural machine translation systems hoping they are good enough at word sense disambiguation since it is a necessary prerequisite for good translation. Despite the fact that this approach failed, we describe it here in order to share knowledge we obtained during this study about weaknesses of currently available machine translation systems and their application to word sense induction.

3.1. Description of the Method

Different senses of a polysemous word in Russian are often translated using different words in English. If a translation system is good enough to produce correct translations, we could find an English word it used to translate a polysemous Russian word and utilize it as a sense identifier. Then we simply group together all examples with the same sense identifier. To implement this approach one needs to decide how to find the word in the translated text corresponding to the polysemous word in the source text. Also, there are often several occurrences of the polysemous word probably in different forms in the source text. Next, we describe how we deal with these issues.

3.2. Discussion of the Results

In the preliminary experiment we compared two machine translation systems available online, namely Google Translate[6] and Yandex Translate[7]. All examples of several words from active-dict and bts-rnc were translated using each system, the annotators were instructed to select the translation of the first occurrence of the target word using system's alignment (both systems highlight phrases in the source text corresponding to the selected phrase in the target text) and use it as a sense identifier for the target word. We did not try automatic alignment methods like fast_align [Dyer et.al., 2013] due to the lack of time. Hence the reported results are a kind of upper bound for machine translation approach to word sense induction.

Figures 4 and **5** show the results of the machine translation based method compared to word2vec weighted average clustering method for several words from active-dict and bts-rnc train sets. Notice that only subset of words from the train sets was annotated, so this average ARI could not be compared to the average ARI on the whole train set **reported in 3.1**. The results of MT systems were relatively bad compared even to the **non-tuned approach from 3.1** (word2vec average with tf-idf weights followed by agglomerative clustering with 2 clusters) and the tuned version with both tf-idf and chi2 weights. To improve the results we tried normalizing translations (yandex-normalized, google-normalized in the figures) using the Porter stemmer from NLTK[8]. This improved average ARI slightly on bts-rnc and worsened it on the active-dict.

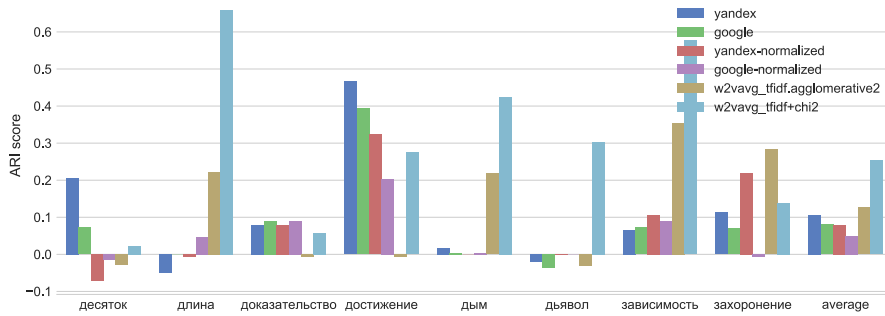


Figure 4. Comparison of machine translation and weighted word2vec average methods on the train set. ARI for several words from the active-dict train set and their average is presented

6 http://translate.google.com, accessed on 21.12.2017–24.12.2017.

7 http://translate.yandex.ru, accessed on 21.12.2017–24.12.2017.

8 http://www.nltk.org

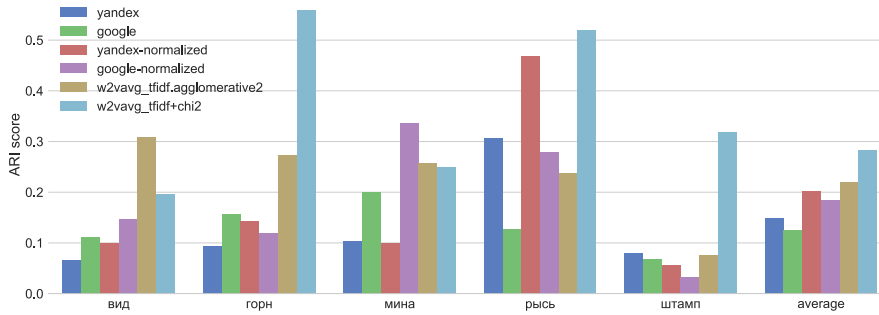


Figure 5. Comparison of machine translation and weighted word2vec average methods on the train set. ARI for several words from the bts-rnc train set and their average is presented

Table 4. A context from the wiki-wiki dataset translated by neural machine translation systems. All occurrences of the word "банка" in the source text should be translated as "jar".

source	трехлитровая **банка** во времена СССР такие банки были популярны для маринованных овощей , овощных и фруктовых соков и так далее . популярность трехлитровых банок объясняется тем , что это самая объемная банка из массово доступных , и это удобно при большом объеме заготовок . в наши дни стеклянные банки продолжают использоваться в быту для домашнего консервирования . подготовка абсолютно целых (без трещин и сколов) стеклянных банок подразумевает не только тщательное мытье внутри и снаружи слабым
Yandex Translate	three-liter **jar** in Soviet times, such banks were popular for pickled vegetables , vegetable and fruit juices and so on . the popularity of three-liter cans is explained by the fact that this is the largest Bank of massively available , and it is convenient with a large volume of blanks . nowadays, glass jars continue to be used in everyday life for home canning . training brand whole (without cracking and chipping) glass jars involves not only a thorough wash inside and outside of the weak
Google Translate	three-liter **bank** during the Soviet Union, such banks were popular for pickled vegetables, vegetable and fruit juices and so on. the popularity of three-liter cans is explained by the fact that this is the most voluminous bank of mass available, and this is convenient for a large volume of blanks. Today glass banks continue to be used in everyday life for home canning. the preparation of absolutely whole (without cracks and chips) glass jars implies not only thorough washing inside and out with the weak

After preliminary experiments on active-dict and bts-rnc we hypothesized that the performance of MT may be better for the wiki-wiki dataset which mostly contains words with coarse-grained senses, so the next experiment was performed on the wiki-wiki test set and the results were submitted to the leaderboard. We noticed that Yandex Translate produced less fluent translations but gave better ARI, so it was chosen for this experiment. Also this time we instructed our annotators to select translations of all occurrences of the target word. To reduce errors each example was annotated by two annotators and the differences (which were very few) were eliminated by the third one. All translation were normalized and the most frequent translation for each example was used as its sense identifier. The resulting submission received 0.8125/0.3957 public/private ARI ranking 4th/12th correspondingly.

To explain the poor performance of MT-based approach we performed error analysis and noticed that translation systems are very inconsistent in their translation of polysemous words. A different occurrence of these words in a single text is often translated differently (cf. **Table 4** for an example). One explanation of this inconsistency is that MT systems are trained on pairs of sentences, not pairs of texts, since currently available machine learning algorithms have problems dealing with long sequences. For instance, Yandex Translate splits input text into sentences and translate each sentence in isolation, hence it cannot take into account context from neighboring sentences[9]. We however, cannot be sure that Google Translate also performs sentence based translation. However, its performance for WSI is no better than that of Yandex.

Despite the great improvement in machine translation quality after switching from phrase based to neural based systems there is still a very large gap between machine and human translation quality and it is very unlikely to disappear in the next few years. For instance, English→French machine translation quality measured by BLEU score on news-test-2014 dataset improved from 37 to 41 points over the last 3 years while the professional human translator quality is estimated as 50 points[10]. Word sense disambiguation is probably one of the biggest challenges which should be solved for machine translation systems in order to outperform humans.

4. Error Analysis

In this section, we analyze the errors made by both approaches described in this paper. Consider the words "*горн*", "*рысь*", and "*штамп*" from the bts-rnc dataset and "*дым*", "*зависимость*" from the active-dict dataset. Confusion matrices for these words are presented in Figure 6, their rows correspond to the true sense identifiers and our interpretation of them (in round brackets) and their columns represent either clusters built by the weighted word2vec average method or translations of the target word by the machine translation based method (the translations were stemmed resulting in absence of several few letters). The cell values are the number of examples (contexts).

[9] This information was provided in personal communication from a Yandex Translate developer.

[10] https://www.eff.org/ai/metrics#Translation

For the word "*штамп*" we could not distinguish the second sense from the first one. The other three senses (stamp, press die and cliche) are distinguished by the weighted average method in majority of examples while the machine translation uses the word "*stamp*" instead of "*cliche*" for the third sense most of the time. For the word "*дым*" (smoke, mist or household in ancient Russia) the machine translation method uses only the word "smoke" while the other method doesn't distinguish the first and the second senses but correctly separates the household sense. The three senses of the word "*зависимость*" (mathematical correlation, political or economic dependence and addiction) can really be translated with the same word "dependence" so the base hypotheses behind machine translation approach to WSI doesn't hold in this case and the approach fails. Conversely, the weighted average method makes very few mistakes for this word.

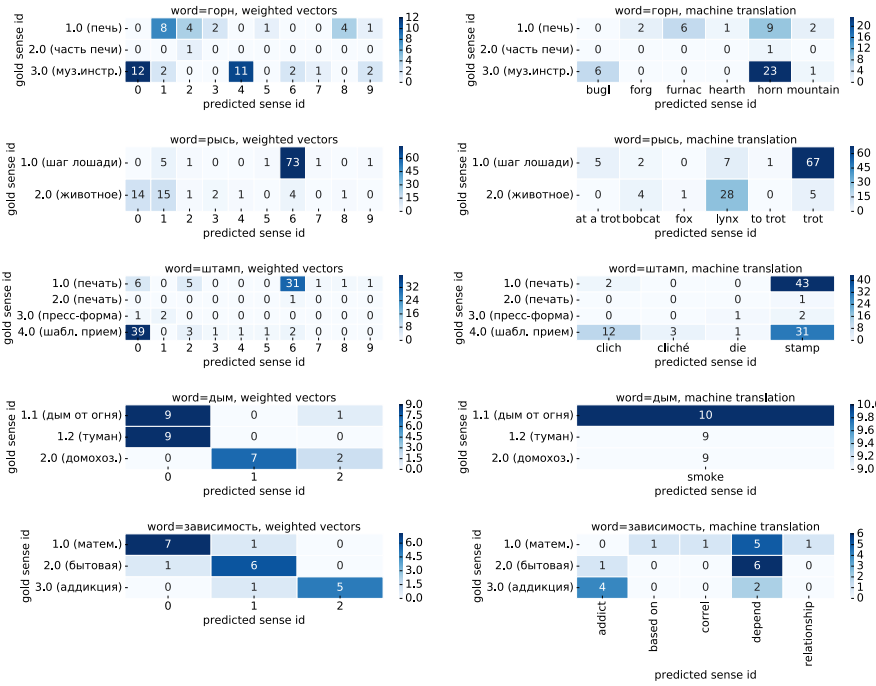


Figure 6. Confusion matrices for weighted word2vec average and machine translation methods

From these examples we can make a conclusion that the machine translation systems sometimes don't work for WSI simply because they correctly use the same word to translate both senses, but more often because they don't translate ambiguous words correctly. Although the weighted word2vec average method shows better results, sometimes it also confuses quite distant senses (stamp and cliche sense of the word "*штамп*", for instance). Also it is unsatisfactory that on two out of three datasets simple clustering algorithms with the same number of clusters for all words work

better than more advanced algorithms which select the number of clusters for each word individually. We plan to investigate the reasons of these problems and to search for a solution in the future work.

5. Conclusion

The paper describes two experiments on word sense induction and disambiguation for the Russian language: a positive and a negative one in terms of their results. The first (successful) one is based on clustering of contexts represented using a weighted average of word embeddings. The second (unsuccessful) is based on the state-of-the-art neural machine translation systems: the translations of ambiguous words into a different language are used as sense labels. Results of the evaluation campaign RUSSE'2018 show that the first approach yields very competitive results, compared to other 18 participating teams, ranking second on two datasets and third on the rest one. Besides, our method substantially outperforms competitive state-of-the-art baselines based on the AdaGram word sense embeddings. Interestingly, despite the expectations, the second approach based on a sophisticated production machine translation systems yielded non-competitive performance.

6. Acknowledgements

Alexander Panchenko was supported by Deutsche Forschungsgemeinschaft (DFG) under the projects "JOIN-T" and "ACQuA".

References

1. *Agirre E., Soroa A.* (2007), Semeval-2007 Task 02: Evaluating Word Sense Induction and Discrimination Systems, Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval '07), Prague, Czech Republic, pp. 7–12.
2. *Alagić D., Šnajder J., Padó S.* (2018): Leveraging Lexical Substitutes for Unsupervised Word Sense Induction. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18).
3. *Bartunov S., Kondrashkin D., Osokin A., Vetrov D.* (2016): Breaking Sticks and Ambiguities with Adaptive Skip-gram. International Conference on Artificial Intelligence and Statistics (AISTATS),
4. *Arefyev N., Panchenko A., Lukanin A., Lesota O., Romanov P.* (2015): Evaluating Three Corpus-Based Semantic Similarity Systems for Russian. In Proceedings of the 21st International Conference on Computational Linguistics and Intellectual Technologies (Dialogue'2015). Moscow, Russia. RGGU
5. *Brendan J. Frey; Delbert Dueck* (2007): Clustering by passing messages between data points" Science. 315 (5814): 972–976
6. *Dyer C., Chahuneau V., Smith N. A.* (2013): A Simple, Fast, and Effective Reparameterization of IBM Model 2. In Proceedings of NAACL-HLT 2013, pages 644–648.

7. *Corrêa Jr E. A., Amancio D. R.* (2018). Word sense induction using word embeddings and community detection in complex networks. arXiv preprint arXiv:1803.08476.
8. *Miller G. A., Beckwith R., Fellbaum C., Gross D., Miller K.* (1990): WordNet: An online lexical database. International Journal of Lexicography, vol. 3, pages 235–244.
9. *Jurgens D., Klapaftis I.* (2013), SemEval-2013 Task 13: Word Sense Induction for Graded and Non-Graded Senses, Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, Georgia, USA, pp. 290–299.
10. *Kobritsov B. P., Lashevskaja O. N., Shemanaeva O. Yu.* (2005), Shallow Rules for Word-Sense Disambiguation in Text Corpora, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2005" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii «Dialog 2005»], Zvenigorod, Russia.
11. *Lopukhin K., Lopukhina A.* (2016), Word Sense Disambiguation for Russian Verbs Using Semantic Vectors and Dictionary Entries, Computational Linguistics and Intellectual Technologies: Papers from the Annual conference "Dialogue" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Po Materialam Ezhegodnoy Mezhdunarodnoy Konferentsii «Dialog»], Moscow, Russia, pp. 393–405.
12. *Lopukhina A., Lopukhin K.* (2016): Word Sense Disambiguation for Russian Verbs Using Semantic Vectors and Dictionary Entries. Computational Linguistics and Intellectual Technologies. Dialogue 2016, pages 393–405.
13. *Loukachevitch N., Chuiko D.* (2007), Thesaurus-based Word Sense Disambiguation [Avtomaticheskoe razreshenie leksicheskoy mnogoznachnosti na baze tezaurusnykh znaniy], Proceedings of the Contest "Internet Mathematics 2007" [Sbornik rabot uchastnikov konkursa «Internet-Matematika 2007»], Yekaterinburg, Russia, pp. 108–117.
14. *Lyashevskaja O., Mitrofanova O.* (2009), Disambiguation of Taxonomy Markers in Context: Russian Nouns, Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009), Odense, Denmark, pp. 111–117.
15. *Manandhar S., Klapaftis I., Dligach D., Pradhan S.* (2010), SemEval-2010 Task 14: Word Sense Induction & Disambiguation, Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval '10), Los Angeles, CA, USA, pp. 63–68.
16. *Mikolov, T., Chen, K., Corrado, G., Dean, J.* (2013), Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR.
17. *Panchenko A., Ustalov D., Arefyev N., Paperno D., Konstantinova N., Loukachevitch N. and Biemann C.* (2016), Human and Machine Judgements about Russian Semantic Relatedness. In Proceedings of the 5th Conference on Analysis of Images, Social Networks, and Texts (AIST'2016). Communications in Computer and Information Science (CCIS). Springer-Verlag Berlin Heidelberg
18. *Panchenko A., Lopukhina A., Ustalov D., Lopukhin K., Leontyev A., Arefyev N., Loukachevitch N.* (2018), RUSSE'2018: A Shared Task on Word Sense Induction and Disambiguation for the Russian Language. In Proceedings of the 24rd International Conference on Computational Linguistics and Intellectual Technologies (Dialogue'2018). May 30—June 2, Moscow, Russia

19. *Panchenko A., Fide M., Ruppert E., Faralli S., Ustalov D., Ponzetto S. P., Biemann C.* (2017), Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguation. In Proceedings of the the Conference on Empirical Methods on Natural Language Processing (EMNLP). Copenhagen, Denmark. Association for Computational Linguistics
20. *Pelevina M., Arefyev N., Biemann C., Panchenko A.* (2016), Making Sense of Word Embeddings. In Proceedings of the 1st Workshop on Representation Learning for NLP co-located with the ACL conference. Berlin, Germany. Association for Computational Linguistics
21. *Ustalov D., Panchenko A., Biemann C.* (2017), Watset: Automatic Induction of Synsets from a Graph of Synonyms, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, pp. 1579–1590.

MORPHOLOGICAL SEGMENTATION WITH SEQUENCE TO SEQUENCE NEURAL NETWORK

Arefyev N. V. (narefjev@cs.msu.su)[1]
Lomonosov Moscow State University, Moscow, Russia;
Samsung Moscow Research Center, Moscow, Russia

Gratsianova T. Y. (tgratsianova@cs.msu.su),
Popov K. P. (kpopov94@ya.ru)[1]
Lomonosov Moscow State University, Moscow, Russia

Morphological segmentation is an important task of natural language processing as it can significantly improve the processing of unfamiliar and rare words in different tasks that involve text data. In this paper we present datasets in English and Russian for learning and evaluating morphological segmentation algorithms, demonstrate the method based on the sequence to sequence neural model and show that the proposed approach shows better results in comparison with other existing methods of morpheme segmentation. We start from an English dataset, which is already available and only minor preprocessing has been made, and then we experiment with the Russian language, where we could not obtain prepared data. So, some more serious preprocessing issues are included. Moreover, we demonstrate how morphological segmentation can improve another natural language processing task—evaluation of words semantic similarity. To achieve this goal, first we try to reproduce the best results of the participants of Russian words semantic similarity competition (RUSSE), which was conducted in Dialogue 2015 conference. Then we show how with the help of smart morpheme segmentation these results can be advanced.

Keywords: morphological segmentation, sequence transduction, sequence to sequence, semantic similarity

[1] These two authors contributed equally to this work

МОРФЕМНАЯ СЕГМЕНТАЦИЯ С ПОМОЩЬЮ SEQUENCE TO SEQUENCE НЕЙРОННОЙ СЕТИ

Арефьев Н. В. (narefjev@cs.msu.su)[2]
Московский Государственный Университет им. Ломоносова, Москва, Россия;
Samsung Moscow Research Center, Москва, Россия

Грацианова Т. Ю. (tgratsianova@cs.msu.su),
Попов К. П. (kpopov94@ya.ru)[2]
Московский Государственный Университет им. Ломоносова, Москва, Россия

1. Introduction

The importance of automatic morphological segmentation lies in the fact that it improves the processing of rare and unknown words, which are common in natural texts. Usually the algorithm designed for solving this task takes words of some language as input, and returns as output the same words, but segmented into morphemes.

Nowadays there are many systems for natural language processing, which are trained on huge amounts of text data. Taking a word for a minimal language unit is a common approach in such algorithms. However, it is known that, for example, in Russian there is an order of magnitude more words than the morphemes. Here by the term *morpheme* we mean the minimal meaningful part of the word. That is why the model that was trained on morphemes instead of words will be much smaller. This fact will allow using such model on devices with limited memory.

Another advantage of using morphemes in natural language processing tasks is the possibility of handling unknown and rare words. For example, if we want to evaluate semantic similarity of two unknown words, we can split them into morphemes and somehow estimate similarity between these morphemes. The fact that morphemes are minimal meaningful parts of words guarantees that such evaluation will be justified.

In this paper we describe the possibility of morphological segmentation with sequence to sequence (seq2seq) model and evaluate results of this approach by different methods. Our main contributions are the following.

1. We adapt sequence to sequence neural network for morphological segmentations and show its superiority over existing models on Russian and English datasets.

[2] Эти два автора внесли одинаковый вклад в эту работу

2. We develop a new dataset for the Russian language for training and evaluation the methods of morphological segmentation. It is described in the section where we present other used datasets.
3. We show that our approach improves the semantic similarity estimation of unknown words.

We compare our method with the existing universal algorithm and with the algorithm developed specially for Russian language, xMorphy[3].

We open sourced our code to facilitate further research in morphological segmentation and it's applications for the Russian language[4].

2. Morphological Segmentation as Sequence Transduction

Sequence to sequence (seq2seq) is a general-purpose neural network architecture for sequence transduction, which is used for tasks such as machine translation, text summarization, conversational modeling and more as described in [Denny Britz et al., 2017]. In this work, we adapt seq2seq, consisting of an encoder and decoder, with an attention mechanism for morphological segmentation. The next three paragraphs briefly describe the encoder, decoder and the attention mechanism.

An encoder reads in «source data», e.g. a sequence of symbols, and produces a vector containing information about this data relevant for the task. The idea is that the representation produced by the encoder can be used by the decoder to generate correct output (solve the task).

A decoder is a generative model that is conditioned on the representation created by the encoder. For example, a Recurrent Neural Network decoder may learn to generate the translation of the encoded sentence into another language.

Instead of encoding the input sequence into a single fixed size representation, the model can, with attention mechanism [Bahdanau et al., 2014], learn how to generate an input representation for each output time step. In other words, the model learns which elements of the input sequence to attend to in order to generate the next element of output sequence, based on the input sequence and what it has produced so far.

For training the model, morphological segmentation task was defined as sequence transduction, that is, the sequence of symbols is being transformed into another sequence of symbols. For this purpose, every word in training datasets was represented as the sequence of its letters, e.g. *б|е|з|о|к|о|н|ы|ъ* (*w|i|n|d|o|w|l|e|s|s*). Additionally, the special symbol "*" was added into target training dataset. This symbol indicated the boundaries between word's segments, e.g. *б|е|з|*|о|к|о|н|*|ы|ъ* (*w|i|n|d|o|w|*|l|e|s|s*).

Hyperparameters, which we used in training process, are described in the **Table 1** (the same hyperparameters were used in every experiment). The values were taken from authors' recommendations for the amount of data which is close to our [Denny B. et al., 2017].

[3] https://github.com/alesapin/XMorphy

[4] https://github.com/kpopov94/morpheme_seq2seq

Table 1. Seq2seq training hyperparameters

Name	Value	Description
Attention hyperparameters		
num_units	256	Hidden state dimension.
Encoder hyperparams		
num_units	256	Size of the LSTM cell in the encoder
dropout_input_keep_prob	0.8	Apply dropout to the (non-recurrent) inputs of each GRU layer using this keep probability.
num_layers	1	Number of GRU layers.
Decoder hyperparameters		
num_units	256	Size of the LSTM cell in the decoder
dropout_input_keep_prob	0.8	Apply dropout to the (non-recurrent) inputs of each GRU layer using this keep probability.
num_layers	2	Number of GRU layers.
Other hyperparameters		
embedding.dim	256	Dimensionality of the embedding layer.

This model is fully described in [Denny Britz et al., 2017] where the schematic picture of its' operation can be found.

3. Related Work

To date, quite a large number of algorithms have been developed for automatic morpheme segmentation. Basically, such tools use approaches based on the principle of maximum likelihood as in [Creutz M. et al., 2004], MAP [Creutz M. et al., 2007], FSA [Goldsmith J. et al., 2004] and CRF[5].

Until 2010, the annual MorphoChallenge competition was held, where different algorithms of morpheme segmentation were compared. In 2010, a program called Morfessor, which is based on the maximum a posteriori estimation principle showed the best results[6]. Later, in 2013, the Morfessor 2.0 was developed and showed much better results than its predecessor improving F-measure from 60% to 80%. The main innovation of this algorithm was the possibility of learning on both labeled and unlabeled data. It was also possible to set hyperparameters for balancing the importance between the labeled and the unlabeled data to several thousands, as the algorithm authors do for the best results. For example, if the annotated corpus was relatively small, it was necessary to increase the beta coefficient, which was responsible for the weight of labeled data during training.

In this paper, we compare our method to Morfessor 2.0, because it shows better quality than Morfessor 1.0 which was the winner in MorphoChallenge 2010, as was mentioned above.

[5] https://github.com/alesapin/XMorphy

[6] http://morpho.aalto.fi/events/morphochallenge2010/comp1-results.shtml

4. Evaluation

There are two main approaches to morphological segmentation evaluation. In direct evaluation the results of an algorithm are compared to gold standard. Indirect evaluation shows the benefits of predicted morphological segmentations on some other task.

4.1. Direct Evaluation

For the direct evaluation we chose Boundary Precision and Recall (BPR)[7] metric, which was used in MorphoChallenge competition since 2005. This choice was mainly motivated by the fact that the authors of Morfessor 2.0 also used it.

As in many other evaluation methods, precision, recall and F-measure for single word are calculated by these formulas:

$$precision = \frac{number\ of\ correct\ boundaries\ found}{total\ number\ of\ boundaries\ found}$$

$$recall = \frac{number\ of\ correct\ boundaries\ found}{total\ number\ of\ correct\ boundaries}$$

$$F - measure = 2 * \frac{precision * recall}{precision + recall}$$

Here the term *boundary* means the borderline between word's segments, which the algorithm succeeds or fails to discover. For example, there is one boundary in the word *nepe*eзд* (*re*location*). If we assume, that algorithm segmented this word as *nep*e*eзд* with two boundaries, we can calculate precision as 0.5, recall as 1.0 and F-measure for this word will be 0.66.

For calculating total precision, recall and F-measure simple average value are taken through them.

4.2. Indirect Evaluation

As a task for indirect evaluation of morphological segmentation, we decided to use the task of Russian words Semantic Similarity Estimation (RUSSE) introduced as shared task in Dialogue 2015 conference [Panchenko A. et al., 2015]. The results of comparison of three algorithms for solving this task are presented in [Arefyev N. V. et al., 2015]. The best results were shown by the word2vec model and that's why we used it in our experiments.

Word2vec is designed for training on large text data for representing the words as relatively low dimensional dense vectors. This approach allows estimating words similarity using dot product between corresponding vectors. The main disadvantage of this model, perhaps, is the absence of predictions about words that were not included in the model and for rare words the estimation will be totally incorrect.

[7] http://morpho.aalto.fi/events/morphochallenge/software/bpr.py

To solve this problem, the method for resolving the absence of out-of-vocabulary words (denoted as OOV) was proposed. If the next word cannot be found in the model during the evaluation process, this method tries to find some part of this word in the model. That is, the search is done by sequential separation of prefixes from word, letter by letter.

For the demonstration and evaluation of morphological segmentation we have improved this approach. Instead of separating leading symbols, the word is segmented into morphemes with the help of seq2seq model and then the resulted segments and their sequential concatenations are searched in the model. For example, let us assume that the word *жертвовательница* (*sacrificer*) is not in the word2vec model and is segmented as *жертв*ова*тель*ниц*а*. The next processing of this word can be shown in the Table 2 (simplified for demonstration):

Table 2. Word processing in morpheme segmentation resolution approach

window	segments	found in model
1	*жертв* *ова* *тель* *ниц* *а*	none
2	*жертвова* *ователь* *тельниц* *ница*	none
3	*жертвователь* *овательниц* *тельница*	*жертвователь*
4	*жертвовательниц* *овательница*	none

The following estimation of semantic similarity will be done for the word *жертвователь* (*sacrifier*). To be short, hereinafter this method will be referenced as MSR (morpheme segmentation resolution).

As the conclusion for this section, it is worth to mention that we are not solving the word embedding task in this work, but we use this task for evaluation of morphological segmentation quality.

4.3. Datasets

For the experiments in this paper we used several different datasets. Some of the data was preprocessed for lowercasing, replacing letters “ё” to “е”, deleting extra symbols and so on. For seq2seq model training data was represented in the form described in Section 2. In the following description of the data we use the term “word type” to denote unique words in the data and the term “token” to denote a particular occurrence of a word:

1. All English text data came from MorphoChallenge 2010 competition[8]:
   - 878,036 unsegmented word types for training
   - 1,000 segmented word types for training
   - 686 segmented word types for testing
2. Lib.rus.ec book collection—424,362 unsegmented words for training.
3. [Tikhonov A. N. 2008]—98,186 segmented word types, that were used in experiments both for training and evaluation[9].
4. Russian Wikipedia—238,052,379 tokens, that were used in RUSSE competition in 2015. We used this corpus for training word2vec model for repeating results from [Arefyev et al., 2015] in Evaluation on RUSSE task section and for experiments with our approach in this task.
5. Datasets from RUSSE competition. They are HJ, RT, AE and AE2 datasets and fully described in [Panchenko A. et al., 2015].

For the training and testing purposes, words from Tikhonov were randomly divided into train and test sets in 3:1 proportion.

5. Results

5.1. BPR on English datasets

For initial evaluation of the proposed method, we decided to compare it with Morfessor on English dataset, which was used in [Sami Virpioja et al., 2013]. We used the same trainsets and the same hyperparameters for training Morfessor as in [Sami Virpioja et al., 2013].

Seq2seq was trained with hyperparameters described in **Section 2**.

We used 1000 segmented words for training every algorithm and we also used 878,036 unsegmented words for Morfessor training.

The difference between original results for Morfessor, presented in [Sami Virpioja et al., 2013], and our reproduced results on the same training dataset can be explained by the different test sets. The authors of Morfessor 2.0 have not mention where it is possible to get the test set they used in experiments in their paper, so we just used the remaining words (i.e. the words have not been used for training) from MorphoChallenge 2010 dataset.

Results are shown in the **Table 3**:

[8] http://morpho.aalto.fi/events/morphochallenge2010/datasets.shtml#download

[9] Russian dataset for seq2seq training and evaluation available at https://drive.google.com/file/d/1DcAVZ4Nv5Xbeua8vnW4SUylnxGqvR5yn/view

Table 3. Results for English dataset

Method	Test data	Precision	Recall	F-measure
Morfessor [Sami Virpioja et al., 2013]	1,000	0.8591	0.8550	0.8571
Morfessor (reproduced)	686	0.8676	0.8530	0.8603
Seq2seq	686	0.9019	0.8716	0.8865

As we can see, seq2seq model shows both better precision and better recall compared to Morfessor 2.0, even though it could not exploit large amount of unsegmented data.

We also compared our approach to MORSE, one of the recently developed algorithms for morphological segmentation. It is fully described in [Tarek Sakakini, 2017]. The authors of MORSE published only one trained model for English, and no source code for training were provided.

Test data were truncated to 539 words, because for 147 words MORSE did not provide predictions, which are equal to their origin words, e.g. *accompanied* turned to *accompany|ed*. The BPR evaluation method that we used does not accept such situations. Results are in the **Table 4**.

Table 4. Morfessor, MORSE and seq2seq comparison

Method	Precision	Recall	F-measure
Morfessor	0.8663	0.8719	0.8691
MORSE	0.9125	0.6785	0.7783
Seq2seq	0.8914	0.8922	0.8918

Seq2seq again showed the best results, and even Morfessor demonstrated its superiority against MORSE. The difference between MORSE and seq2seq is more than 0.1.

5.2. BPR on Russian datasets

Models for Russian words were trained on Tikhonov dictionary [Tikhonov A. N., 2008] as described in **Section 4.3**. For Morfessor training we used unsegmented lib.rus.ec corpus, which is also described in **Section 4.3**.

Morfessor was trained with hyperparameters from previous section and seq2seq was trained with parameters, which were described in **Section 2**.

We also compared seq2seq model and Morfessor 2.0 to xMorphy[10] tool, which is the only tool we found for morphological segmentation of words in Russian. Unfortunately, the authors do not supply code for training so we could not train it on our data.

We used 73,639 segmented words for seq2seq and Morfessor training and also 424,362 unsegmented words for Morfessor. For test we took the remaining part of Tikhonov dictionary which was 24,547 words.

[10] https://github.com/alesapin/XMorphy

Table 5. Results for Russian dataset

Method	Precision	Recall	F-measure
xMorphy	0.7797	0.6589	0.7143
Morfessor	0.9127	0.8904	0.9014
Seq2seq	0.9407	0.9383	0.9395

We can see from results that seq2seq gives 4% and 22% better F-measure than Morfessor and xMorphy respectively.

Considering the fact that BPR applies only the segmentation boundaries by evaluation algorithm, there are counts for different types of errors in the Table 6 for the demonstration purposes. While counting errors, the segmentation was considered correct only if it fully coincided with the gold standard, otherwise it was considered incorrect.

Table 6. Comparison of count of errors done by two algorithms

	Morfessor correct	Morfessor incorrect
seq2seq correct	13,806 words	5,396 words
seq2seq incorrect	2,695 words	2,650 words

As we can see, seq2seq shows better results than Morfessor. And if we take a closer look on separate words, we can notice that errors by seq2seq were made with relatively difficult words. There are some random examples in Table 7:

Table 7. Random examples of complicated segmentations

Word	Seq2seq result	Gold standard segmentation
венчиковый	вен*чик*ов*ый	вен*ч*ик*ов*ый
забытье	за*бы*ть*е	забы*ть*е
статичный	статич*н*ый	стат*ич*н*ый
расточать	рас*точ*а*ть	расточ*а*ть
скрыться	с*кры*ть*ся	скры*ть*ся
шилоклювка	шил*о*клюв*к*а	ши*л*о*клюв*к*а
подержаться	подерж*а*ть*ся	по*держ*а*ть*ся

5.3. Evaluation on RUSSE task

In the evaluation on RUSSE task we used the same seq2seq model as we did in previous section, and benchmarks described in [Arefyev et al., 2015].

Four semantic similarity approach were done (first two are the same as in the [Arefyev et al., 2015]):

1. Without out-of-vocabulary optimization.
2. With out-of-vocabulary optimization (OOV).
3. With MSR optimization.
4. With MSR and OOV (where MSR failed) optimization.

Results are in the Table 8:

Table 8. RUSSE evaluation results

Method	HJ	RT	AE	AE2
No optimization [Arefyev et al., 2015]	0.53200	0.73100	0.88100	0.91400
No optimization (reproduced)	0.52964	0.73563	0.88310	0.91253
OOV (reproduced)	0.55314	0.81217	0.91381	0.91909
MSR	0.56845	0.82738	0.91448	0.91941
MSR + OOV	0.56845	0.82849	0.91507	0.92039

A little difference in first two rows of the table can be explained by differences in the text preprocessing before word2vec training.

As we can see from the results, the method with morphological segmentation shows better results for every evaluation dataset. The Table 9 contains the words for which MSR was able to improve word similarity evaluation (without optimization the value would be zero). Unknown words are marked by "?" signs.

Table 9. Examples of MSR improvements

First word	Second word	First word + MSR	Second word + MSR	Words similarity measure
осмысление	?осмысливание?	осмысление	мысл	0.62
авиасообщение	?авиауслуга?	авиасообще-ние	авиа	0.48
аудио	?аудиопродукция?	аудио	аудио	1.00
?ахвахец?	?ахвахска?	ахвах	ахвах	1.00
сатана	?люциферов?	сатана	люцифер	0.61

6. Conclusion and Further Work

After all of experiments, the proposed model demonstrates a sustainable superiority over Morfessor 2.0. We also proved that the morphemic segmentation can improve the results of word semantic similarity estimation.

A relatively small improvement on the RUSSE problem can be developed by improving the algorithm for finding concatenations of morphemes in the model due to some limitations, for example, you can try to limit the size of segments for which search will be made and others not to be considered, and also focus only on options with root morphemes.

As we said earlier, the representation of natural language as units in the form of vectors for solving natural language processing problems is quite common. If we take words as units, several problems immediately arise:

1. The model produces a large number of vectors, which directly affects the size of the model.
2. For rare words, vectors are unrepresentative, that is, the vector representation of a rare word does not carry in itself useful data and does not reflect reality.

3. Since natural language is a constantly expanding system, it is inevitable that many words of natural language will not enter the model anyway, no matter how large is the amount of data for training.

On the other hand, letters could be taken as language units. Then it would take only, for example, 33 vectors for the Russian language, since there are only 33 letters in it. This would significantly reduce the volume of the model, but the practical use of this approach would be extremely small, since letters do not carry semantic information. It is known that morphemes are the minimal meaningful parts of words. The language has much less morphemes than words. This fact allow this representation to solve the problem of a large volume of the model. The problem of rare words will be greatly simplified due to the fact that even rare words consist of morphemes, which, at least some of them, will be recognized by the model. The disadvantage of this approach is an increase in ambiguity.

References

1. *Arefyev N. V., Panchenko A. I., Lukanin A. V., Lesota O. O., Romanov P. V.* (2015), Evaluating Three Corpus-based Semantic Similarity Systems for Russian, available at: http://www.dialog-21.ru/media/1119/arefyevnvetal.pdf
2. *Bahdanau D., Cho K., Bengio Y.* (2014), Neural Machine Translation by Jointly Learning to Align and Translate, available at http://www.cl.uni-heidelberg.de/courses/ws14/deepl/BahdanauETAL14.pdf
3. *Creutz M., Lagus K.* (2004), Induction of a simple morphology for highly-inflecting languages. In Proceedings of th7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON). Barcelona, Spain. 43–51.
4. *Creutz M., Lagus, K.* (2007), Unsupervised models for morpheme segmentation and morphology learning. ACM Trans. Speech Lang. Process. 4, 1, Article 3.
5. *Denny Britz, Anna Goldie, Minh-Thang Luong, Quoc Le* (2017), Massive Exploration of Neural Machine Translation Architectures, available at: arxiv.org/pdf/1703.03906.pdf
6. *Goldsmith J., Hu Y.* (2004), From signatures to finite state automata. Midwest Computational Linguistics Colloquium, available at: https://newtraell.cs.uchicago.edu/files/tr_authentic/TR-2005-05.pdf
7. *Panchenko A., Loukachevitch N. V., Ustalov D., Paperno D., Meyer C. M., Konstantinova N.* (2015), RUSSE: The First Workshop on Russian Semantic Similarity, In proceeding of Dialogue 2015 conference, available at: http://www.dialog-21.ru/media/1123/panchenkoaetal.pdf
8. *Sami Virpioja, Peter Smit, Stig-ArneGrönroos, Mikko Kurimo* (2013), Morfessor2.0: Python Implementation and Extensions for Morfessor Baseline, Aalto University publication series, available at: https://aaltodoc.aalto.fi/bitstream/handle/123456789/11836/isbn9789526055015.pdf
9. *Tarek Sakakini, Suma Bhat, Pramod Viswanath* (2017), MORSE: Semantically Driven MORpheme SEgment-er, available at: https://arxiv.org/pdf/1702.02212.pdf
10. *Tikhonov A. N.* (2008), Morpheme-spelling dictionary of the Russian language [Morfemno-orfograficheskij slovar' russkogo yazyka], ACT, Moscow, Russia.

FRAMEWORK FOR RUSSIAN PLAGIARISM DETECTION USING SENTENCE EMBEDDING SIMILARITY AND NEGATIVE SAMPLING

Belyy A. V. (anton.belyy@gmail.com)[1,2],
Dubova M. A. (marina.dubova.97@gmail.com)[3]

[1]ITMO University, Saint Petersburg, Russia;
[2]B Tochka Bank QIWI Bank (JSC), Yekaterinburg, Russia;
[3]Saint Petersburg State University, Saint Petersburg, Russia

In this paper, we propose a new approach for advanced plagiarism detection in Russian language. It is based on a classifier, dealing with two different types of sentence similarity measures: token set similarity and cosine similarity between sentence embeddings (based on pre-trained RusVectōrēs, unsupervised fastText, and supervised StarSpace models). The diversity of feature space makes it possible to detect different types of plagiarism, starting from simple copy&paste cases and ending with complex manual paraphrases. The proposed approach implies an ability to focus on the particular plagiarism type identification, allowing to train a universal model at the same time. The method shows great results on detection of different types of plagiarism and outperforms the previous approach.

Keywords: plagiarism detection, sentence similarity, word embeddings, negative sampling

МЕТОД ПОИСКА ТЕКСТОВЫХ ЗАИМСТВОВАНИЙ НА ОСНОВЕ ВЕКТОРНОЙ БЛИЗОСТИ ПРЕДЛОЖЕНИЙ С ОТБОРОМ КАНДИДАТОВ

Белый А. В. (anton.belyy@gmail.com)[1,2],
Дубова М. А. (marina.dubova.97@gmail.com)[3]

[1]Университет ИТМО, Санкт-Петербург, Россия;
[2]Ф Точка Банк КИВИ Банк (АО), Екатеринбург, Россия;
[3]Санкт-Петербургский Государственный Университет, Санкт-Петербург, Россия

1. Introduction

Plagiarism is a perennial and pervasive problem in research and educational institutions [Maurer et al., 2006]. Nowadays, it is becoming even more widespread as a number of free Internet resources grows. Failures to appropriately acknowledge original sources discount a reward system in science [Resnik et al., 2005] and obstruct scientific search of the relevant literature. In education, plagiarism spoils an assessment process [Larkham and Manns, 2002] and makes it difficult to guarantee a correspondence between certificates and real knowledge. Thus, it impairs science and education systems and turns out to be one of the most serious problems in these fields.

In Russia, the research on advanced plagiarism detection methods is only starting to gain popularity. In 2016–2017, the first competition, PlagEvalRus-2017, was held. Organizers provided participants with a large and diverse dataset in Russian [Smirnov et al., 2017], [Sochenkov et al., 2017], and made an appropriate evaluation possible [Smirnov et al., 2017].

It this paper, we propose a new solution to the Russian plagiarism detection problem. Our approach employs the diversity of Natural Language Processing methods to deal with plagiarism cases of different complexity. It achieves great results and outperforms the previous method in detecting plagiarism of all types.

2. Task and dataset

2.1. Task

Scientists split the problem of plagiarism detection into two main parts: source retrieval and text alignment (for a review of subtasks, see [Alzahrani et al., 2012]). Source retrieval is a search of all possible sources for documents suspected to borrow information, whereas text alignment is a search of particular contiguous passages of text borrowed from a given source. In our work, we consider only the second task and address an extrinsic plagiarism detection. It means that we already have suspicious documents with source candidates and try to find particular fragments, matching in suspected and source documents.

2.2. Dataset

There are plenty of datasets for plagiarism text alignment in English, which were employed at the special PAN competitions held annually from 2009 to 2014. The Russian dataset is a new one (for a detailed description, see Smirnov et al., 2017, Sochenkov et al., 2017), nevertheless, it was developed according to the worked-out "gold-standards" of English plagiarism detection datasets. It consists of three main types of plagiarism:

- Automatically generated copy&paste plagiarism (copy&paste)
- Automatically generated paraphrased plagiarism (paraphrased)
- Manually generated paraphrased plagiarism (manual)

3. Related work

The main contributions to the plagiarism text alignment were made at the PAN annual competitions, and the latest winning approach is described in [Sanchez-Perez et al., 2014]. In this method, authors use TF-IDF weighted vectors of sentences and cosine similarities and the Dice coefficients between them for comparison. Afterwards, scientists usually focused their work on the particular type of plagiarism (in most of cases, the manual type) and achieved significant improvements by applying genetic algorithms [Vani and Gupta, 2017]; [Sanchez-Perez et al., 2017], word embedding models [Brlek et al., 2016], and topic modeling methods [Le et al., 2016].

However, only one method is adapted to Russian language. This algorithm [Zubarev and Sochenkov, 2017] was developed in accordance with a specificity of Russian language and PlagEvalRus-2017 dataset. Authors applied semantic and syntactic parsing to compute textual similarity score, used to detect plagiarism cases. This method was developed and evaluated at the first competition on advanced plagiarism detection methods in Russia.

4. Method

4.1. Preprocessing

4.1.1. Text splitting

We split each document into sentences using *nltk* library. Afterwards, each sentence is tokenized using simple regular expression and lemmatized using *mystem* library to reduce the size of the vocabulary. To account for homographs (words with the same spelling, but different pronunciation and meaning, like добро́$^{\text{noun}}$, добро́$^{\text{part}}$ and до́бро$^{\text{adv}}$ in Russian), we also store part of speech (PoS) tags of each lemma estimated using *mystem*. Sentences of short length (less than three tokens) are merged together with adjacent longer sentences to reduce granularity of predictions. After this step, we obtain a list of sentences, which are represented as sequences of lemmas with PoS tags.

4.1.2. Sentence matching

For text classifier training set (see 4.2.1) we need to construct sentence-to-sentence matchings from text-to-text matchings originally provided in the dataset. In the most common case, we have a short text of n sentences from a suspicious document which we want to match with a short text of m sentences from a source document and obtain $f(m,n)$ sentence-to-sentence pairs. In our work we experimented with two simple types of matching:

1. parallel matching: each sentence i from the suspicious text is matched with the corresponding sentence i from the source text, producing $min\,(m,n)$ matchings
2. pairwise matching: each sentence from the suspicious document is matched with all sentences from the source text, producing $m * n$ matchings

Each matching makes sense for different corpora, depending on the distribution of sentence lengths from the particular dataset. For PlagEvalRus-2017 dataset, we use parallel matching in copy&paste and paraphrased datasets and use pairwise matching in manual datasets.

4.2. Model

4.2.1. Classifier

A key component of our method is a classifier, predicting a fact of plagiarism for a given pair of sentences. In experimental phase, we tested and compared three classifiers:

- Logistic Regression (with L2 regularization and C=1.0)
- Random Forest (with 10 trees)
- Bayesian classifier, where a posteriori probabilities $p(y|x)$ are obtained using non-parametric kernel density estimation (KDE) described in [O'Brien et al, 2016].

4.2.2. Feature space

To obtain the utility of our approach for very different types of plagiarism, we decided to combine simple similarity measures along with more complex and modern ones, such as cosine similarity of sentence embeddings. Each pair of sentences (u,v) is therefore represented as a sequence of different similarity measures: $(s_1(u,v), s_2(u,v), \ldots, s_n(u,v))$, where each measure $s_i(u,v)$ is normalized to have value range $[0;1]$ and $s_i(u,u) = 1$.

4.2.2.1 Token similarity

Given tokens of a pair of sentences, denoted as u and v, we calculate how much common vocabulary do these two sentences share, using the following measures:

- $LeftInclusion(u,v) = \frac{|u \cap v|}{|u|}$
- $RightInclusion(u,v) = \frac{|u \cap v|}{|v|}$

A combination of these measures can confidently detect a copy&paste or a light paraphrase, however, it performs poorer in harder cases, for which we need to measure sentence similarity in a different way.

4.2.2.2 Sentence embeddings

Given a pair of sentences u and v, represented as sequences of tokens, we would like to train a mapping (or embedding) f from sentence to vector, such that if sentences u and v are semantically related, then the angle between vectors $f(u)$ and $f(v)$ is close to 0 (or, equivalently, cosine distance of $f(u)$ and $f(v)$ is close to 1):

$$\frac{< f(u), f(v) >}{|f(u)| \cdot |f(v)|} \approx 1$$

Our approach is based on combining several kinds of word embeddings, each optimized for different objective and therefore capturing different aspects of sentence similarity:

- RusVectōrēs [Kutuzov and Kuzmenko, 2017] pre-trained model, which was trained on Russian National Corpus and Russian Wikipedia (600 million tokens, resulting in 392,000 unique word embeddings), thus allowing to introduce more general similarities into the model.
- fastText [Joulin et al., 2016] unsupervised model, which was trained on texts from PlagEvalRus-2017 datasets (12 million tokens, resulting in 184,000 unique word embeddings). One useful feature of the fastText model is that it uses internally character N-grams rather than tokens, which is helpful in the presence of rare / out-of-vocabulary words.
- StarSpace [Wu et al., 2017] supervised model, which was also trained on PlagEvalRus-2017 corpus (resulting in 97,000 unique word embeddings). This recently introduced framework has many use-case scenarios, one which is supervised similarity learning between sentences.

We obtain sentence embeddings by averaging embeddings of individual words multiplied by their IDF weights estimated on each dataset separately [Ferrero et al., 2017]:

$$f(sent) = \frac{1}{|sent|} \sum_{w \in sent} f(w) \cdot idf(w)$$

4.3. Negative sampling

Data points in the original dataset are coordinates of textual fragments in source and suspicious documents, which correspond to each positive plagiarism case. The resting pairs of short texts can be used as negative cases to train the classifier. However, this approach has two important disadvantages: high imbalance of classes in the training set (as reflected in Table 1) and time consumption. This forced us to try **Random-*N*** negative sampling. In this approach, *N* negative examples per every positive are chosen randomly from the whole set of negative examples at the train time.

Table 1. Label distribution in different parts of PlagEvalRus-2017 dataset

	Plagiarism	Non-plagiarism	All	Plagiarism, %
Copy&paste	95,122	101,796,105	101,891,227	0.09%
Paraphrased	122,867	113,941,646	114,064,513	0.11%
Manual	18,366	16,856,074	16,874,440	0.11%
Manual2	4,241	2,648,296	2,652,537	0.16%

4.4. Granularity reduction

We reformulate the original text alignment problem to binary classification over pairs of sentences. After that, we need to merge adjacent plagiarised sentences into contiguous passages to reduce granularity (section 6.1) of predictions. We propose a simple algorithm that runs in linear time with respect to the number of sentence-to-sentence detections:

Algorithm 1. Granularity reduction
Parameters:

- gap_{susp}, gap_{src}—maximum distance between adjacent sentences from suspicious and source documents;

Variables:

- D—list of all detections in a pair of documents, ordered by suspicious sentence id;
- S—mapping from detection to a component id;

```
1. for each detection(susp_i,src_j) ∈ D:
2.     if(susp_i,src_j):
3.         S[susp_i,src_j]= new component id
4.     for i in 0..gap_susp:
5.         for j in 0..gap_src:
6.             if (susp_i+i,src_j+j) ∈ D:
7.                 S[susp_i+i,src_j+j]=S[susp_i,src_j]
```

After that, we merge detections with the same component id into a single prediction. Parameters gap_{susp}, gap_{src} are the maximum look-ahead distance of the algorithm and need to be tuned for each dataset.

5. Model selection and parameter tuning

5.1. Feature space

Feature space size put some limitations on the set of possible classifiers and account for the time and memory consumption. Considering these problems, we reduced the initial feature space from 5 to 2 measures by grouping and averaging metrics of the similar nature:

- TokenSim = ½ (LeftInclusion + RightInclusion)
- SentEmbed = ⅓ (RusVectōrēs + StarSpace + fastText)

We will use manually paraphrased part of PlagEvalRus-2017 dataset to compare classifiers in the rest of this section.

As captured in Table 2, performances of LR and RF did not change significantly. Moreover, the reduction to 2 features makes our results more interpretable and susceptible for further analysis.

Table 2. Train results for plagiarism detection on the manually paraphrased plagiarism

	Granularity	Macro			Micro		
		Precision	Recall	Plagdet	Precision	Recall	Plagdet
PAN Baseline	1.2145	**0.8899**	0.4471	0.5189	**0.9393**	0.5419	0.5992
Log Regression 2D	**1.0001**	0.7246	0.8558	0.7848	0.8173	0.8759	0.8455
Log Regression 5D	**1.0001**	0.7300	0.8537	0.7870	0.8176	0.8733	0.8445
Random Forest 2D	1.0003	0.7809	0.9340	**0.8504**	0.8347	0.9450	0.8862
Random Forest 5D	1.0004	0.7690	**0.9487**	0.8492	0.8263	**0.9566**	**0.8865**

5.2. Negative sampling

We compare sampling methods to see how they affect train time and quality of sentence similarity prediction task. For this comparison, we took 20% sentence pairs as a hold-out set and used the rest 80% for training. A quality measure is classifier's ROC-AUC on hold-out pairs.

Table 3. Quality of sentence similarity on 2D features with undersampling

	No sampling	Random-1	Random-10	Random-100
Log Regression	**0.9785**	**0.9785**	**0.9785**	**0.9785**
Random Forest	0.9322	0.9734	0.9669	0.9533
Bayesian KDE	0.9755	0.9762	0.9757	0.9758

Table 4. Training time for sentence similarity on 2D features with undersampling

	No sampling	Random-1	Random-10	Random-100
Log Regression	**22.45**	**0.06**	**0.21**	**2.27**
Random Forest	383.70	0.24	1.76	26.10
Bayesian KDE	40.47	2.16	2.63	6.28

We see that Logistic Regression (LR) achieves the best quality in the smallest amount of time. Bayesian KDE is on par with LR but takes more time to train. Random Forest demonstrates the worst quality and training time. Importantly, all classifiers show their best performance in Random-1 sampling, which justifies the use of undersampling in plagiarism detection tasks.

However, we should still beware class imbalance in test time. Below are decision boundaries for LR-2D classifier trained with different sampling parameters:

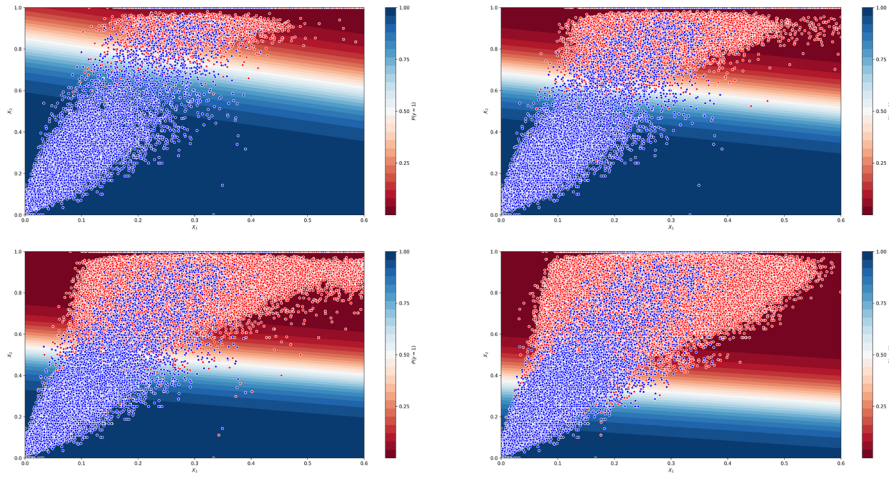


Figure 1. LR-2D trained on Random-N for N = 1, 10, 100 and no sampling

As the number of negative samples grows, the slope of decision line remains roughly the same, but its margin (y-intercept) gets closer to zero. In order to achieve good quality on test set, we have to tune classifier's margin b on non-sampled labeled hold-out set $S = (Sx, Sy)$ using some function Q which measures how close are predictions of trained classifier $a(Sx)$ to true labels Sy:

$$b = argmax_b Q(Sy, [a(Sx) - b > 0])$$

For Q we propose to use either classifier F_1-score or train Plagdet score (defined in section 6.1). The former takes less time to compute, but the latter is a more accurate estimate of test Plagdet score. Below are results for tuning b with classifier F_1-score:

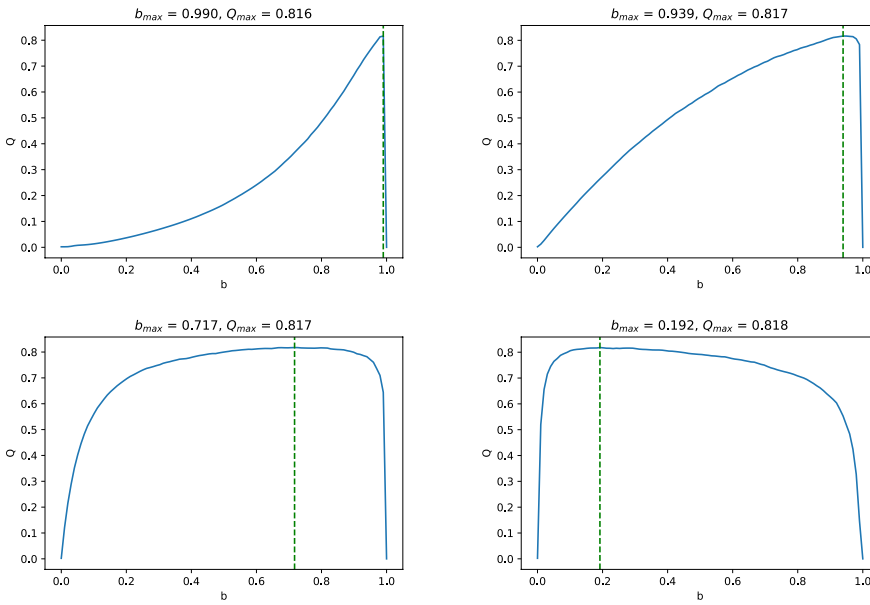


Figure 2. Tuning margin for LR-2D with Random-N for N = 1, 10, 100 and no sampling

5.3. Parameter tuning

For the whole model we need to tune three hyperparameters: gap_{susp}, gap_{src} (section 4.4) and margin b (section 5.2) and choose N for Random-N. To compute test Plagdet, we selected the best (in terms of ROC-AUC) classifier from section 5.2—LR-2D with Random-1 sampling. We trained four distinct models (three on different parts of the dataset and one on the whole dataset) and selected their hyperparameters using grid search on a linear and logarithmic scale to maximize train Plagdet:

Table 5. Hyperparameters for models

	Random-N	Granularity gap_{susp}	Granularity gap_{src}	Margin b
Copy&paste	1	10	10	$1-10^{-6}$
Paraphrased	1	10	10	$1-10^{-5}$
Manual	1	0	0	$1-10^{-2}$
All parts	1	10	10	$1-2{\cdot}10^{-5}$

6. Results

6.1. Evaluation metrics

Standard evaluation metrics for text alignment task are [Potthast et al., 2010]:

- Macro-averaged and micro-averaged **precision** and **recall**.

$$prec_{macro}(S,R) = \frac{1}{|R|}\sum_{r\in R}\frac{|\cup_{s\in S}(s\sqcap r)|}{|r|} \qquad prec_{micro}(S,R) = \frac{|\cup_{(s,r)\in(S\times R)}(s\sqcap r)|}{|\cup_{r\in R} r|}$$

$$rec_{macro}(S,R) = \frac{1}{|S|}\sum_{s\in S}\frac{|\cup_{r\in R}(s\sqcap r)|}{|s|} \qquad rec_{micro}(S,R) = \frac{|\cup_{(s,r)\in(S\times R)}(s\sqcap r)|}{|\cup_{s\in S} s|}$$

- **Granularity** measure. It represents a mean number of plagiarism detections per a single positive case. Ideally, an algorithm should find only one case of plagiarism for each fragment.

$$gran(S,R) = \frac{1}{|S_R|}\sum_{s\in S_R}|R_S|$$

- **Plagdet** score: a combination of precision, recall, and granularity, considering the importance of each measure. It is considered to be the most representative and valid metric for evaluation of plagiarism detection methods.

$$plagdet(S,R) = \frac{F_1}{log_2(1+gran(S,R))},$$

where F_1 is a harmonic mean of macro- or micro-averaged precision and recall.

6.2. Testing

We trained LR-2D model with Random-1 sampling on different parts of the dataset and came up with four different models: trained on all, copy&paste, paraphrased and manual types of plagiarism respectively. Afterwards, we evaluated them on hold-out test sets from PlagEvalRus-2017 corpora (for evaluation details, see Smirnov et al., 2017). Models were tested on certain types of plagiarism and on the whole dataset, obtained results are presented in Tables 1–4.

Table 6. Test results for copy&paste plagiarism

	Granularity	Macro			Micro		
		Precision	Recall	Plagdet	Precision	Recall	Plagdet
PAN Baseline	**1.0046**	0.7240	0.9101	0.8038	0.9615	**0.9943**	**0.9744**
zubarev17.1	1.5084	**0.9496**	0.6427	0.5778	**0.9828**	0.8217	0.6746
zubarev17.2	1.4660	0.9320	0.7013	0.6146	0.9776	0.8588	0.7022
Belyy: all types	1.0202	0.8168	0.8790	0.8347	0.9131	0.9717	0.9280
Belyy: copy&paste	1.0066	0.8711	0.8269	**0.8444**	0.9447	0.9396	0.9377

	Granularity	Macro			Micro		
		Precision	Recall	Plagdet	Precision	Recall	Plagdet
Belyy: paraphrased	1.0147	0.8297	0.8718	0.8413	0.9210	0.9665	0.9333
Belyy: manual	1.1336	0.3652	**0.9322**	0.4800	0.5994	0.9935	0.6839

Table 7. Test results for paraphrased plagiarism

	Granularity	Macro			Micro		
		Precision	Recall	Plagdet	Precision	Recall	Plagdet
PAN Baseline	3.4639	0.9051	0.6895	0.3626	0.9710	0.8334	0.4156
zubarev17.1	1.5404	**0.9604**	0.6730	0.5884	**0.9875**	0.8219	0.6670
zubarev17.2	1.4834	0.9473	0.7340	0.6303	0.9812	0.8650	0.7006
Belyy: all types	1.0111	0.8535	0.8788	0.8591	0.9186	0.9649	0.9337
Belyy: copy&paste	**1.0039**	0.9169	0.7760	0.8382	0.9532	0.9113	0.9292
Belyy: paraphrased	1.0074	0.8694	0.8668	**0.8635**	0.9286	0.9579	**0.9380**
Belyy: manual	1.1378	0.4259	**0.9460**	0.5359	0.6179	**0.9936**	0.6951

Table 8. Test results for manual plagiarism

	Granularity	Macro			Micro		
		Precision	Recall	Plagdet	Precision	Recall	Plagdet
PAN Baseline	1.1414	0.8332	0.0554	0.0946	**0.8960**	0.0761	0.1277
zubarev17.1	1.0015	0.8068	0.3409	0.4788	0.8845	0.3815	0.5325
zubarev17.2	1.0016	0.6250	0.4715	**0.5369**	0.8208	0.5312	0.6443
Belyy: all types	1.0000	0.8054	0.2824	0.4181	0.7910	0.2993	0.4343
Belyy: copy&paste	1.0000	0.8384	0.0505	0.0953	0.8714	0.0539	0.1015
Belyy: paraphrased	1.0000	**0.8500**	0.2138	0.3417	0.8208	0.2223	0.3499
Belyy: manual	1.0038	0.2412	**0.8700**	0.3767	0.6030	**0.8912**	**0.7173**

Table 9. Test results for all types of plagiarism

	Granularity	Macro			Micro		
		Precision	Recall	Plagdet	Precision	Recall	Plagdet
PAN Baseline	1.9953	0.8525	0.3366	0.3049	0.9637	0.6893	0.5078
zubarev17.1	1.3028	**0.9129**	0.4605	0.5087	**0.9693**	0.7043	0.6780
zubarev17.2	1.2417	0.8158	0.5644	0.5729	0.9460	0.7737	0.7309
Belyy: all types	1.0104	0.8312	0.5075	**0.6256**	0.9015	0.7898	**0.8357**
Belyy: copy&paste	**1.0048**	0.8910	0.3340	0.4842	0.9472	0.6928	0.7975
Belyy: paraphrased	1.0080	0.8492	0.4612	0.5944	0.9158	0.7648	0.8288
Belyy: manual	1.0564	0.3506	**0.8961**	0.4846	0.6073	**0.9662**	0.7170

Basing on macro-averaged plagdet, our algorithm shows the best results on 3 of 4 tests: copy&paste, paraphrased, and all types of plagiarism. However, the previous approach [Zubarev and Sochenkov, 2017] is still the best for the manual type. As for micro-averaged plagdet, our method turned out to be the best for paraphrased, manual, and all types of plagiarism. The baseline for copy&paste detection is very high, however, 3 of 4 models got very close to this result.

Models trained on the specific type of plagiarism show the best performance on their targets. Moreover, the model trained on the whole dataset not only achieved the best result on the whole test set (all types) but got very close to the best results on copy&paste and paraphrased parts. Also, it achieved the best micro-averaged plagdet for manual plagiarism.

However, for manual plagiarism the result turned out to be skewed (recall is high, but precision is low; or recall is low, but precision is high) for both current and previous approach, meaning that on this part of of the dataset even specific models fail to generalise appropriately. Additional work with detailed error analysis on manual plagiarism is required.

7. Conclusion and further directions

In this paper, we propose a new approach to Russian plagiarism text alignment and show how the diverse methods of Natural Language Processing can be successfully applied in one framework. The diversity of employed metrics perfectly matches heterogeneity of target problem. We use simple and standard metrics of textual similarity in combination with complex and modern ones (such as supervised sentence embeddings). This set allows to deal with simple and hard plagiarism cases at the same time.

A trained classifier unites the diverse metrics into a comprehensive structure. It allows teaching models focused on a specific type of plagiarism, keeping on the opportunity to train a universal one at the same time. This enlarges the possible set of problems to deal with and makes our approach flexible.

Finally, the algorithm shows great performance on all types of plagiarism in Russian language, significantly outperforming previous methods. This is the primary marker of its universality and high quality.

The main direction for improvement seems to be an enlargement and transformation of feature space to get better results on manual plagiarism detection. It could be achieved by taking into account syntactic similarity, adding word N-grams to token set similarities, and using more advanced approaches for classification, such as such CNN or LSTM models. Sentence embeddings feature set can be extended to include vectors from sent2vec [Pagliardini et al., 2017] and doc2vec [Le et al., 2014] models, which build distributed representations of sentences and short texts rather than words. Another serious improvement could be a reduction of hyperparameter space, such as classifier margin . Additional study on negative sampling in plagiarism detection may yield fruitful results here.

References

1. *Brlek A., Franjic P., and Uzelac N.* (2016), Plagiarism detection using word2vec model, Text Analysis and Retrieval 2016 Course Project Reports, pp. 4–7.
2. *Ferrero, J., Agnes, F., Besacier, L., & Schwab, D.* (2017), Using Word Embedding for Cross-Language Plagiarism Detection, arXiv preprint, arXiv, 1702.03082.
3. *Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T.* (2016), Fasttext.zip: Compressing text classification models, arXiv preprint, arXiv:1612.03651.
4. *Kutuzov, A., & Kuzmenko, E.* (2016), WebVectors: a toolkit for building web interfaces for vector semantic models, In International Conference on Analysis of Images, Social Networks and Texts, pp. 155–161.
5. *Larkham, P. J., & Manns, S.* (2002), Plagiarism and its treatment in higher education. Journal of Further and Higher Education, Vol. 26(4), pp. 339–349.
6. *Le H. T., Pham L. M., Nguyen D. D., Nguyen S. V., and Nguyen A. N.* (2016), Semantic text alignment based on topic modeling, In Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2016 IEEE RIVF International Conference on. IEEE, pp. 67–72.
7. *Maurer H., Kappe F., and Zaka B.* (2006), Plagiarism-a survey, J. UCS, Vol. 12(8), pp. 1050–1084.
8. *O'Brien, T. A., Kashinath, K., Cavanaugh, N. R., Collins, W. D., & O'Brien, J. P.* (2016), A fast and objective multidimensional kernel density estimation method: fastKDE, Computational Statistics & Data Analysis, Vol. 101, pp. 148–160.
9. *Potthast M., Stein B., Barron-Cede A., and Rosso P.* (2010), An evaluation framework for plagiarism detection, In Proceedings of the 23rd international conference on computational linguistics: Posters, Association for Computational Linguistics, pp. 997–1005.
10. *Resnik D. B.* (2005), The ethics of science: An introduction, Routledge.
11. *Sanchez-Perez M., Gelbukh A., Sidorov G., and Gomez-Adorno H.* (2017), Plagiarism detection with genetic-based parameter tuning. International Journal of Pattern Recognition and Artificial Intelligence, pp. 1860006.
12. *Smirnov I., Kuznetsova R., Kopotev M., Khazov A., Lyashevskaya O., Ivanova L., Kutuzov A.* (2017), Evaluation Tracks on Plagiarism Detection Algorithms for the Russian Language. International Conference "Dialogue 2017" Proceedings, Vol. 2, pp. 271–285.
13. *Sochenkov I. V., Zubarev D. V., Smirnov I. V.* (2017), The ParaPlag: Russian Dataset for Paraphrased Plagiarism Detection. International Conference "Dialogue 2017" Proceedings, Vol. 1, pp. 284–294.
14. *Sanchez-Perez M., Sidorov G., and Gelbukh A.* (2014), A winning approach to text alignment for text reuse detection at PAN 2014, In CLEF (Working Notes), pp. 1004–1011.
15. *Vani K., Gupta D.* (2017), Detection of idea plagiarism using syntax–semantic concept extractions with genetic algorithm, Expert Systems with Applications, Vol. 73, pp. 11–26.
16. *Wu, L., Fisch, A., Chopra, S., Adams, K., Bordes, A., & Weston, J.* (2017), StarSpace: Embed All The Things!, arXiv preprint, arXiv:1709.03856.

17. *Zubarev D. V., Sochenkov I. V.* (2017), Paraphrased Plagiarism Detection Using Sentence Similarity, International Conference "Dialogue 2017" Proceedings, Vol. 2, pp. 408–418.
18. *Pagliardini M., Gupta P., Jaggi M.* (2017), Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features, arXiv preprint, arXiv:1703.02507.
19. *Le Q., & Mikolov T.* (2014), Distributed representations of sentences and documents. In International Conference on Machine Learning (pp. 1188–1196).

QUALITY EVALUATION AND IMPROVEMENT FOR HIERARCHICAL TOPIC MODELING

Belyy A. V. (anton.belyy@gmail.com)

ITMO University, Saint Petersburg, Russia;
B Tochka Bank QIWI Bank (JSC), Yekaterinburg, Russia

Seleznova M. S. (maria.selezniova@phystech.edu),
Sholokhov A. K. (ak.sholokhov@gmail.com),
Vorontsov K. V. (vokov@forecsys.ru)

Moscow Institute of Physics and Technology (State University),
Moscow, Russia

Generic topics of large-scale document collections can often be divided into more specific subtopics. Topic hierarchies provide a model for such topic relation structure. These models can be especially useful for exploratory search systems. Various approaches to building hierarchical topic models have been proposed so far. However, there is no agreement on a standard approach, largely due to the lack of quality metrics to compare existing models. To bridge this gap we propose automated evaluation metrics which measure the quality of topic-subtopic relations (edges) of a topic hierarchy. We compare automated evaluations with human assessment to validate the proposed metrics. Finally, we show how the proposed metrics can be used to control and to improve the quality of existing hierarchical models.

Key words: topic modeling; topic hierarchies; quality metrics; coherence; word embeddings

ОЦЕНКА И УЛУЧШЕНИЕ КАЧЕСТВА ИЕРАРХИЧЕСКИХ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ

Белый А. В. (anton.belyy@gmail.com)

Университет ИТМО, Санкт-Петербург, Россия;
Ф Точка Банк КИВИ Банк (АО), Екатеринбург, Россия

Селезнева М. С. (maria.selezniova@phystech.edu),
Шолохов А. К. (ak.sholokhov@gmail.com),
Воронцов К. В. (vokov@forecsys.ru)

Московский физико-технический институт
(государственный университет), Москва, Россия

1. Introduction

Topic modeling is a branch of unsupervised machine learning widely used to summarize large unlabeled text corpora. A probabilistic topic model extracts latent probabilities of words appearing in each topic and topics appearing in each document, uncovering vectors of probability distributions that represent documents.

For the purposes of creating a representation of a text collection that helps users to navigate through the collection smoothly, topics can be arranged into a hierarchy. Generic topics of each parent level are thus divided into more specific subtopics of its child level. Such representation allows users to constrict the set of documents they are interested in gradually going down the topic hierarchy.

Various approaches to topic hierarchy learning have been proposed in recent years, such as LDA [1], hPAM [2] and hARTM [3]. However, there is still no agreement on the common approach. The main problem resides in difficulties of topic hierarchies comparison. Since there is no common topic hierarchy quality metrics, it is currently impossible to compare different approaches rigorously.

A quality metric for hierarchical topic models should measure both interpretability of topics on each hierarchy level and quality of pairs of topics that are linked with parent-child relations in the hierarchy. There are common ways to measure topics quality widely used in the field, such as topic coherence [4]. Also, various topic quality metrics based on word embeddings have been proposed recently [5, 6]. However, to the best of our knowledge, "parent-child" relations quality has not been explored so far. In this paper, we propose metrics for quality of the hierarchy edges which represent such relations.

We use BigARTM—an open source library for topic modeling of large collections—in our experiments.

2. Hierarchical Topic Models

Let D denote a document collection. A vocabulary W is a set of tokens (e.g. words, tags, links, etc.) that appear in the collection. We assume that the collection contains topics from a finite set T. Then, each document $d \in D$ can be described with its probability distribution $(t|d)$ over the topics $t \in T$ (i.e. $p(t|d)$ is a vector of probabilities for each topic to appear in the document d). On the other hand, each topic $t \in T$ is described with its probability distribution $(w|t)$ over the tokens $w \in W$.

Given a collection D, we can extract estimators of a probability distribution $p(w|d)$ of its tokens over its documents as n_{dw}/n_d, where n_{dw} is a number of times the token w appears in the document d, $n_d = \Sigma_{w \in W} n_{dw}$ is a number of words in d. However, we cannot directly estimate $(w|t)$ or $p(t|d)$ as t is a latent (hidden) variable. As described, for example, in [12], extracting those distributions can be formulated as a matrix factorization problem $F=\Phi\Theta$, where $F = \{p(w|d)\}_{W \times D}$ is the given matrix of $p(w|d)$ estimators, $\Phi = \{p\,(w \mid t)\}_{W \times T}$ $\Theta = \{p(t|d)\}_{T \times D}$ and are the matrices of model parameters that we are aimed to find. As shown in [12], the problem can be solved through EM-algorithm application.

A hierarchical topic model (HTM) comprises several flat (described above) topic models that form hierarchy levels. Each (l+1)-th level has more topics than the

l-th one for the topics to get more specific down the hierarchy. HTM also includes edges that represent "parent-child" relations between the topics of the neighboring levels. As shown in [3], the problem of building such a hierarchy level by level can be solved through adding a new matrix $\Psi^l = \{p(t|a)\}_{T \times A}$ that represents probabilities for topics $t \in T^{l+1}$ (a set of topics of the $(l+1)$-th level) to be subtopics of the previous level topics $a \in T^l$. That gives a matrix factorization problem $\Phi^{l+1} = \Phi^l \Psi^l$ for each level [3].

3. Motivation

Fig. 1. Start screen of the exploratory search system Rysearch

Since its inception, topic modeling has been successfully applied for visualizing and navigating through large scientific corpora [8, 9, 10]. flexibility for such visualizations by allowing a user to go down a hierarchy to more specific topics or go up to more general ones. HTMs are particularly promising as a technology for creating an exploratory search engine [11], which allows exploring an area of knowledge related to a user's query rather than looking for the exact query.

Our distant goal was to create such an engine. We have been working on a prototype, which currently indexes Russian popular science websites and blogs and aggregates their content into a hierarchical topic model. In our work, we have faced two major problems:

- **heterogeneity of sources**, which means that our sources usually differ in size and comprise different sets of topics,
- **absence of evaluation metrics for HTMs** for HTMs, which slows down model design as each HTM needs to be evaluated manually before it can be deployed into production environment.

In this paper we propose several metrics for automated model evaluation, thus tackling the second problem. We also share some insights on how these metrics can be used to improve quality of already built hierarchical model.

To facilitate comparison of metrics, we built two hierarchical models, **concat** and **heterogeneous**, which we use throughout the paper for explanations. The first one is intentionally worse in subjective quality than the latter: it has been trained on a simple concatenation of all the sources into a single one, disregarding the inequality of their sizes and structure. The latter model was built using a method we proposed that includes several stages:

1. Build a topic model of a 'base' collection—a collection that allows to build an interpretable model that includes all the topics we want to aggregate—to create an initial estimate for the hierarchy.
2. Rank documents of a collection we want to add to the hierarchy according to their similarity to the base collection. The most similar documents should be ranked first.
3. Add the documents that appeared to be on the top of the ranking list to the base collection in a quantity not exceeding 10% of the base collection size. Build a topic model of the extended collection using the matrix $\Phi_0{}^1$ from the first stage (that contains estimates for the first level topics of the base collection) to initialize a new matrix Φ^1 of the first hierarchy level.
4. Repeat the third step until all the documents from new collection are added to the model. Each time the collection from the previous iteration of the method is referred to as a base collection, so its size increases and we can add more documents on each step.

In our experiments we used Postnauka.ru as a base collection and Habrahabr.ru as an added collection. Postnauka.ru contains all the major topics present in popular science content and its' articles have manually adjusted tags that allow building a model of high interpretability. On the other hand, Habrahabr.ru is focused on IT and contains a lot of irrelevant content such as news and advertisements. Also, Habrahabr.ru collection is much bigger than Postnauka.ru (see the section 5.1 for the detailed datasets description). To rank the documents we used a regressor that measured how similar a given document is to typical popular science articles. As our ranking method is collection-specific we mention some other ideas about ways to perform ranking in the discussion section.

4. Proposed metrics for hierarchies

Proposed in [4], the topic coherence is a classical measure of a topic quality as well as flat topic models' interpretability in general. In particular, one can estimate quality of a model as a whole by taking the average of topics' coherences. However, a hierarchical topic model consists not only of its levels, but also of relations between topics from the neighboring levels, whereas the average coherence of the model's levels takes these dependencies completely out of consideration. Hence, the average coherence fractionally depicts the quality of a hierarchical model. This section is aimed

to bridge the gap by proposing several quality measures for the "parent—child" relations between topics in a hierarchical model.

Linguistic similarity based metrics. We extend the classical flat coherence from [4] to hierarchical coherence to capture either syntagmatic or paradigmatic relatedness of parent and child topics' top tokens [7]. Let us define $w^{(t)}$ as the i-th top token of some topic t. Then $v^{(t)}$ will be the vector corresponding to this top token in some Vector Space Model (VSM). In our experiments we used the pre-trained VSM RusVectōrēs [13], which was trained on Russian National Corpus and Russian Wikipedia (600 million tokens, resulting in 392,000 unique word embeddings). $D(w_1,w_2)$ is a number of documents in some corpus (in our experiments we use Postnauka corpus to calculate cooccurrences, although in general it is more preferable to use big external corpora such as Wikipedia or Twitter) where words w_1 and w_2 have occurred together at least once. $D(w)$ is a document frequency of word w calculated for the same corpus. Then we define our metrics as:

- EmbedSim: $$\frac{1}{C}\sum_{i=1}^{n}\sum_{j=1}^{n} < v_i^{(a)}, v_j^{(t)} > [w_i^{(a)} \neq w_j^{(t)}]$$
- CoocSim: $$\frac{1}{C}\sum_{i=1}^{n}\sum_{j=1}^{n} ln\frac{D(w_i^{(a)}, w_j^{(t)}) + \varepsilon}{D(w_j^{(t)})}[w_i^{(a)} \neq w_j^{(t)}]$$

where $$C = \sum_{i=1}^{n}\sum_{j=1}^{n}[w_i^{(a)} \neq w_j^{(t)}]$$

is a number of word pairs excluding pairs of identical words. We denote the topic of the parent level as t and the topic of the child level as a ("ancestor") here, n is the number of considered top tokens for each topic.

Probabilistic similarity based metrics. We can compare parent and child topics as probability distributions. Two standard similarity measures for distributions P and Q are Hellinger distance and Kullback-Leibler divergence. The first one is a bounded metric and can be interpreted as distance between two topics in some space. The second is an unbounded asymmetric measure and can be interpreted as "how much information will be lost if we substitute parent topic P with some child topic Q".

- HellingerSim: $$1 - \frac{1}{\sqrt{2}}||\sqrt{p(w|t)} - \sqrt{p(w|a)}||_2$$
- KLSim: $$-D_{KL}(p(w|a)||p(w|t))$$

To understand how these metrics work, let us consider an example. We are given 6 "parent-child" pairs of topics that were assessed by humans. Three of them are labeled as **'good'** (there is a semantic similarity between parent and child), other three are labeled as **'bad'** (little or no similarity). On the fig. 2 one can see these pairs on the right, along with their scores given by the EmbedSim metric. The higher the score, the more confident the metric is. On the left there is a distribution of all the edges from an assessment task described in the following section. Y-coordinates

of points are assigned according to metric score, and colors are set by the assessment experts. One can see that there are much more pairs marked as 'bad' than the ones marked as 'good'. As we expect the hierarchy to be sparse (each parent topic has only a few suitable subtopics), this observation corresponds to our expectations. It is also clear from the figure that 'bad' pairs have lower average metric score than the 'good' ones. It means that the EmbedSim metric scores 'good' pairs with higher values and, therefore, correlates with the assessors opinion.

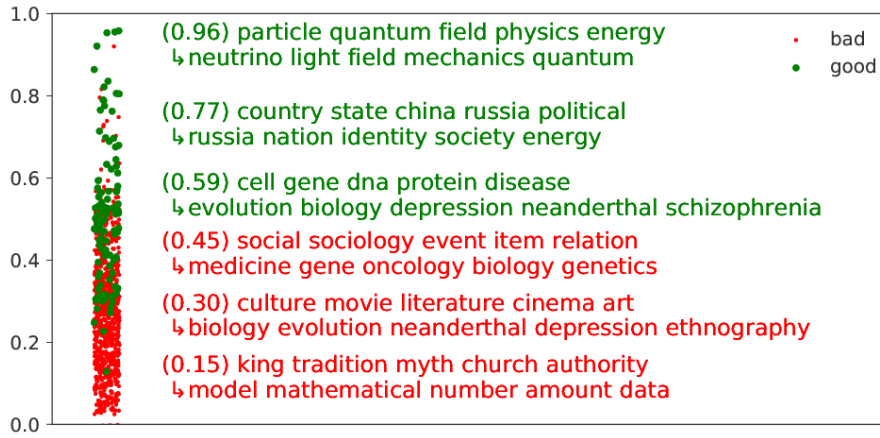


Fig. 2. Examples of topic-subtopic pairs from the assessment task (section 5) scored with the EmbedSim metric as hierarchy edges. Each topic or subtopic is represented by its 5 top tokens. The column on the left shows all the assessed pairs as dots on their EmbedSim score scale

5. Expert opinions of edges quality

5.1. Datasets and models

To construct "parent-child" topic pairs for human annotation, we trained three two-level hierarchical topic models on three datasets:

- **Postnauka.ru**, a popular science website with edited articles on a wide spectrum of topics, focusing on humanities,
- **Habrahabr.ru** and **Geektimes.ru**, social blogging platforms specializing in Computer science, engineering and IT entrepreneurship,
- **Elementy.ru**, a popular science website with a particular focus on life sciences.

Table 1. Datasets' descriptions

Dataset	Number of documents	Unique words	Unique tags	Parent topics	Child topics
Postnauka	2,976	43,196	1,799	20	58
Habrahabr	81,076	588,400	77,102	6	15
Elementy	2,017	30,352	-	9	25

The collections consist of text documents. Dictionary sizes for each collection are listed in the "Unique words" column. Postnauka and Habrahabr collections are also manually tagged by their authors or editors (each article can have multiple tags), the numbers of tags are listed in the "Unique tags" column.

5.2. Task statement

The following question was asked for experts: "given two pairs of topics, T_1 and T_2, decide whether one is a subtopic of another". Possible answers were: "T_1 is a subtopic of T_2", "T_2 is a subtopic of T_1" and "These topics are not related". Topic t was denoted by 10 top words from its probability distribution $(w|t)$.

After the experiment was finished, the first two answers were grouped to denote a single answer "These topics are somehow related" as it was often difficult for assessors to distinguish between a parent and a child given their top words.

5.3. Quality control

To ensure quality control, only those workers who completed training were allowed to enter the assessment task. Experts could have skipped some tasks if they were not sure, but those who were skipping tasks too often were banned from participating for a day.

5.4. Results

Overall, 68 trusted workers participated in our study, each contributed around 100 assessed topical pairs. Assessment of one pair of topics, given their 10 top words, took around 5 seconds for each participants on average. Each topic pair was evaluated by at least five different experts, which gave us 6750 expert annotations for 1350 unique pairs (edges).

Our participants were mainly Russian and Ukrainian nationals, with age varying from 21 to 64 years.

Table 2. Inter-assessor agreement

Agreed assessors	Edge count	Edge percentage
3	374	27.7%
4	468	34.7%
5	508	37.6%

For each pair of topics, we calculate how many assessors made the same verdict (that the topics from the pair are related or that they are not). For 5 assessors per pair, there is always a majority decision, but it can be reached by either 3, 4, or 5 assessors. In the second and the third column we show the quantity and the percentage of the edges with the number of agreed assessors from the first column.

6. Comparison of metrics values and expert annotations

If many people think that there is a "topic—subtopic" relation between two particular topics in a model, a good metric should give a high score for such pair of topics. In this case we say that metric "approximates assessors opinion". Moreover, we want that metric to keep an order on the model edges consistent with this statement: the more people agreed that the relation presents—the higher the metric score should be.

In order to prove that the proposed metric holds this constraint, consider the following classification problem. Let us call "the assessors' judgment" the fact that 4 or 5 assessors agreed on the same verdict (that an edge exists or does not exist in a hierarchy). If it holds, then assessors' judgment on this edge is equal to 1, and −1 otherwise. Let the edges of a hierarchical model be the objects: the positive and negative classes consist of the edges with a positive and a negative assessors' opinion respectively. Let the classifier based on the metric be the following:

$$(t_1,t_2)=\text{sign}(\rho(t_1,t_2)-w)$$

where t_1 and t_2 are the topics from parent and child level of the model respectively, ρ is one of the proposed metrics and w is a margin of the classifier. Having it written in this form, we can calculate ROC AUC for each classifier and estimate the quality of each metric: better approximators are expected to have better scores.

Table 3. ROC AUC scores for the proposed metrics.

Metric	Score
EmbedSim	0.878
CoocSim	0.815
KLSim	0.790
HellingerSim	0.766

The table 3 presents ROC AUC score for each classifier. One can see that the best classification quality was demonstrated by the classifier based on the Embeddings metric (AUC = 0.878). The other metrics demonstrated moderate yet acceptable consistency with the assessors' opinion: AUC values lied evenly above 0.75. For better understanding of this result one can see the fig. 3. For each graph the red line is a density distribution of the metric value for bad edges, and the green one is the same for good edges. The better some vertical line divides bad edges from good ones—the better the metric is. In further experiments we use the EmbedSim metric, as this metrics demonstrated the best consistency with assessors' judgment.

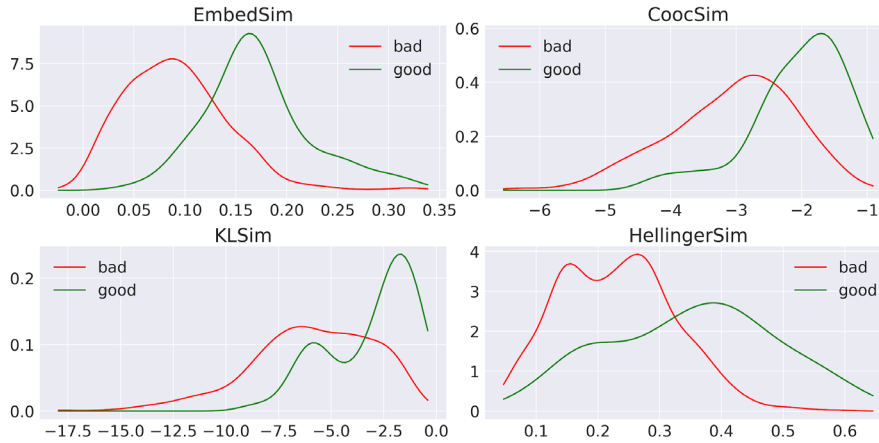


Fig. 3. Distribution of scores for 'bad' and 'good' topical edges

As the EmbedSim metric gives the highest AUC score we use it in all the following experiments.

7. Quality of hierarchical models

The goal now is to combine the edges metric into some construction, which would be a representative quality measure for a hierarchy as a whole.

Normalization: Hereafter we work with a normalized matrix Ψ^{norm} as the following:

$$\psi_{ta}{}^{norm} = \frac{\psi_{ta} - min_t \psi_{ta}}{max_t \psi_{ta} - min_t \psi_{ta}}$$

It allows to apply shared topic-agnostic threshold and to rank all values of Ψ matrix on the same scale.

7.1. Averaging quality

In the spirit of [4] where the average topics coherence was used as a model quality measure, let us consider the average edge quality as quality measure for our hierarchy. The particular hierarchy configuration depends on the chosen threshold for Ψ^{norm}, which determines what probability $(t|a)$ is sufficient to include an edge connecting t and a into the hierarchy. Therefore different thresholds lead to different values of a quality measure. However, for bad models this value seems to be almost evenly lower than for good ones. The fig. 4 illustrates this effect: one can see that the heterogeneous model had a higher score than the less elaborated concat model no matter what threshold was set. Hence it was enough to set the same threshold for all models if one wants to compare them with our measure. However, this measure lacks the interpretability of its value (Y-coordinate of curves on the figure).

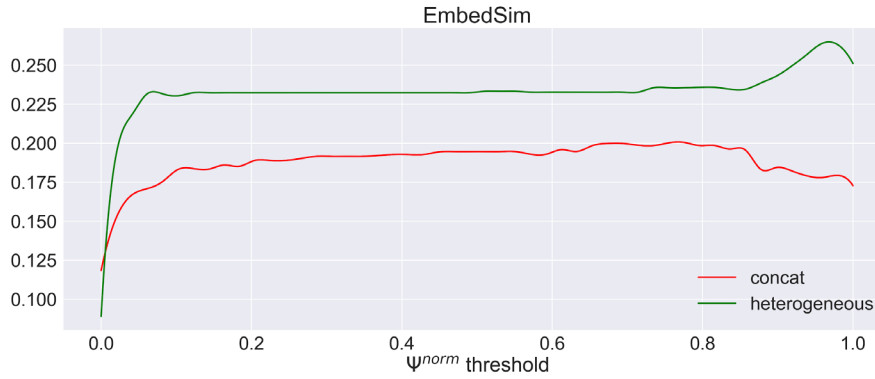


Fig. 4. Averaging quality metrics for EmbedSim

7.2. Ranking quality

Given a collection D, we can extract estimators of a probability distribution $p(w|d)$ of its tokens over its documents as n_{dw}/n_d, where n_{dw} is a number of times the token w appears in the document d, $n_d = \Sigma_{w \in W} n_{dw}$ is a number of words in d. However, we cannot directly estimate $(w|t)$ or $p(t|d)$ as t is a latent (hidden) variable. As described, for example, in [12], extracting those distributions can be formulated as a matrix factorization problem $F=\Phi\Theta$, where $F=[p(w|d)]_{W \times D}$ is the given matrix of $p(w|d)$ estimators, $\Phi=[p(w|t)]_{W \times T}$ $\Theta=[p(t|d)]_{T \times D}$ and are the matrices of model parameters that we are aimed to find. As shown in [12], the problem can be solved through EM-algorithm application.

Another approach to form a quality measure with an interpretable value is to consider the process of establishing a hierarchy as a ranking process. Consider that we have built a model i.e. we have matrices Ψ^{norm}, Θ and Φ for each level. It would be natural to accept only the most meaningful edges according to a human's point of view. As our edge metrics turned out to be good approximators of the assessors' judgment, we can choose only k edges with the top scores of some fixed metrics. If our model is "good", then top-k scored edges (let us call them "the request") should match with top-k maximal elements of the Ψ^{norm} matrix (let us call them "the response"). The difference between the request and the response for each k was measured by common ranking metrics, such as:

- Average Precision (AP@k)—described in [14].
- Inverse Defect Pairs (Inverse DP@k)—the inverse value of the number of pairs that appear in the wrong order (i.e. are reversed) in the response.
- Normalized Discounted Cumulative Gain (nDCG@k)—described in [14].

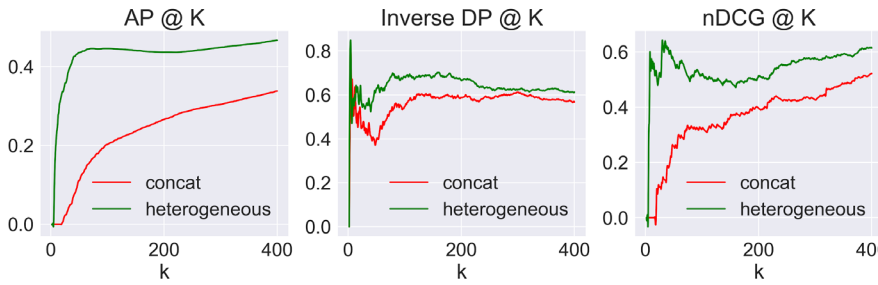


Fig. 5. The ranking quality metrics for the EmbedSim edge metric. The considered models (concat and heterogeneous) are described in section 3

The fig. 5 shows that in all cases the ranking quality scores are higher for better model. One may interpret this result as the following: if a model is "good", than its top-k edges should match the metric's top-k edges precisely enough, no matter what k was set. According to the fig. 4 it holds for all ranking metrics, but the biggest gap was given by the Average Precision. Hence, if one wants to compare quality of two different hierarchies, the advice may be the following:

- Take Embedding similarity (EmbedSim) as the edge metric and plot the Average Precision@k graph. The better model will be the one having better score(s) at the desirable value(s) of k.

There is also a notable advantage of ranking approach over the averaging approach: it allows to choose the optimal number of edges in the model, which we will discuss in the following section.

8. Applications

Using the proposed edge metrics we managed to improve the quality of models significantly in our own project mentioned in section 3. The following chapter briefly describes our experience and results.

8.1. Validating the model in an automated way

Let us suppose that we have a service that continuously aggregates information from various sources into one heterogeneous HTM. It implies that the model mutates over time: new topics and edges appear as the new content arrives. If one does not control the process, the model may degrade over time. To avoid this, we have been using a "good edges to all edges" ratio with automatic notification about model's degradation, so that there is no need in human assessment of the model. The fig. 6 shows some examples of "good", "moderate" and "bad" edges according to the EmbedSim metric.

```
topic_1: file server application user data
    ↳ design internet server file electronics
    ↳ file browser query service web
topic_5: star universe galaxy hole black
    ↳ menu map download mass black
    ↳ astrophysics space gravitation telescope sun
    ↳ space astrophysics sun galaxy gravitation
    ↳ space galaxy telescope astrophysics sun
topic_6: cell gene dna organism protein
    ↳ linux species test result arise
    ↳ evolution bioinformatics paleontology cell zoology
    ↳ dna evolution cell genomics rna
    ↳ evolution ethology cell ecosystem adaptation
    ↳ buffer information open newsletter meditation
    ↳ cell mutation protein dna os
topic_8: brain child neuron memory emotion
    ↳ brain psychology memory neuron thinking
topic_11: country state political authority politics
    ↳ politics state ussr political science authority
```

Fig. 6. Examples of "good" (green), "bad" (red) and "moderate" (pink) edges according to the EmbedSim metric. Each topic or subtopic is represented by its 5 top-tokens.

8.2. Improving the quality of already built models

Another example was an improvement of a previously built model which was too large to be rebuild from scratch. We can't usually change the topics in such a situation, but we can change the hierarchical relations between them.

```
topic_17: server use network data user
    ↳ statistics ecology server ethology network
    ↳ use data statistics the server
topic_5: star planet galaxy bot item
    ↳ astrophysics galaxy planet telescope sun
    ↳ astrophysics sun planet space mass
topic_1: cell organism gene brain animal
    ↳ medicine evolution cell energy geology
    ↳ education school geology usa money
    ↳ cell dna evolution medicine biomedicine
    ↳ work cell evolution species allow
    ↳ medicine evolution cell paleontology immunity
    ↳ cell evolution species give brain
    ↳ medicine evolution physiology energy biomedicine
topic_18: appliance memory brain solution work
    ↳ brain advertising memory neuron cancer
topic_8: company social country society money
    ↳ society state political science england economy
```

```
topic_17: server use network data user
    ↳ use data statistics the server
    ↳ code use function example error
    ↳ install allow information work user
    ↳ file use create data query
    ↳ model method element example code
    ↳ result use point element value
topic_5: star planet galaxy bot item
    ↳ space cosmic particle surface orbit
    ↳ astrophysics sun planet space mass
topic_1: cell organism gene brain animal
    ↳ medicine evolution cell paleontology immunity
    ↳ medicine evolution cell energy geology
    ↳ cell dna evolution medicine biomedicine
topic_18: appliance memory brain solution work
topic_8: company social country society money
    ↳ society culture europe country market
```

Fig. 7. Hierarchical relations improvement with ranking approach edge selection

The fig. 7 (the left side) demonstrates a subset of the parent topics of the concat model with their child topics. According to our method, we plotted an inverse DP@k graph for the EmbedSim metric of the edges of this model (see section 7.2), found its maximum (in our case it was at k = 100) and built a new hierarchy that

contained only the top-k of the edges. The fig. 7 (the right side) demonstrates how quality of the same model increased without rebuilding the model itself. One can see that the new hierarchy looks more consistent and elaborated in comparison with the previous one.

9. Results

In this article we proposed several automated metrics for "parent-child" relations of a topic hierarchy. We showed that the EmbedSim metric based on word embeddings reaches significant consistency with the assessors' judgment on whether the connection between topics exists or not. Other metrics demonstrated moderate yet acceptable consistency and can also be used in conjunction with EmbedSim.

We also proposed two approaches for measuring quality of a hierarchy as a whole. Using metrics of edges' quality we examined averaging and ranking approach to build an aggregated quality measure, and showed that better models reach higher scores in comparison with less elaborated models.

Finally, we demonstrated several applications of the metrics for models' hierarchical relations improvement. For instance, the proposed ranking approach can be used for choosing the optimal set of edges to be included into a hierarchy, as shown in the section 8.

Our work extends existing quality metrics from flat topic models to hierarchical ones which, to the best of our knowledge, hasn't been done before.

10. Discussion

One of the possible extensions of this work is to integrate quality evaluation into the process of building hierarchical models. It can be done in various ways. One option is to build an ARTM regularizer [12], hence the process of constructing a model will try to maximize a certain quality measure during training.

In the section 3 we mention that we used a collection-specific regressor to rank new documents before adding them to the base collection. To make the method applicable to an arbitrary collection we need to replace the regressor with some general approach learned automatically directly from the given collections. One of the possible solutions is, for example, ranking new documents according to their average tf-idf distance to the base collection. Another possible approach is finding $p(t|d)$ vectors of the new documents in the model defined by $\Phi_0{}^1$ matrix from the previous method iteration (i.e. the base collection Φ^1) and, then, measuring their KL divergence with the uniform distribution. The lower its KL divergence is the less the distribution of a given new document over the old topics resembles the uniform distribution. If the distribution is close to uniform it means that the document doesn't fit well into the existing model and should be ranked lower and vice versa.

References

1. *Blei D. M., Griths T., Jordan Michael I., Tenenbaum J.,* (2003), Hierarchical topic models and the nested chinese restaurant process.
2. *Mimno David M., 0010 Wei Li, McCallum Andrew.,* (2007), Mixtures of hierarchical topics with pachinko allocation, ICML / Eds. Zoubin Ghahramani. Vol. 227. P. 633–640. URL: http: //doi.acm.org/10.1145/1273496.1273576.
3. *Chirkova N. A., Vorontsov K. V.,* (2016), Additive regularization for hierarchical multimodal topic modeling, Journal of Machine Learning and Data Analysis [Mashynnoe obuchenie i analiz dannyh], Vol.2 № 2 P. 187–201.
4. *Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A.,* (2011), Optimizing semantic coherence in topic models, In Proceedings of the conference on empirical methods in natural language processing (pp. 262–272). Association for Computational Linguistics.
5. *Fang, A., Macdonald, C., Ounis, I., & Habel, P.,* (2016), Using word embedding to evaluate the coherence of topics from twitter data. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (pp. 1057–1060). ACM.
6. *Nikolenko, S. I.,* (2016), Topic quality metrics based on distributed word representations. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (pp. 1029–1032). ACM.
7. *Schütze, H., & Pedersen, J.,* (1993), A vector model for syntagmatic and paradigmatic relatedness. In Proceedings of the 9th Annual Conference of the UW Centre for the New OED and Text Research(pp. 104–113).
8. *Chaney, A. J. B., & Blei, D. M.,* (2012), Visualizing Topic Models. In ICWSM.
9. *Chuang, J., Manning, C. D., & Heer, J.* (2012), Termite: Visualization techniques for assessing textual topic models. In Proceedings of the international working conference on advanced visual interfaces (pp. 74–77). ACM.
10. *Blei, D. M., & Lafferty, J. D.,* (2007), A correlated topic model of science. The Annals of Applied Statistics, 17–35.
11. *Marchionini, G.,* (2006), Exploratory search: from finding to understanding. Communications of the ACM, 49(4), 41–46.
12. *Vorontsov, K., Frei, O., Apishev, M., Romov, P., Suvorova, M., & Yanina, A.,* (2015), Non-Bayesian additive regularization for multimodal topic modeling of large collections. In Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications (pp. 29–37). ACM.
13. *Kutuzov, A., & Kuzmenko, E.* (2016), WebVectors: a toolkit for building web interfaces for vector semantic models. In International Conference on Analysis of Images, Social Networks and Texts, pp. 155–161.
14. *Hang Li* (2011), A Short Introduction to Learning to Rank. IEICE Trans. Inf. & Syst., Vol.E94–D, No.10, pp. 1854–1862.

SEMANTIC ANALYSIS WITH INFERENCE: HIGH SPOTS OF THE FOOTBALL MATCH

Boguslavsky I. M. (bogus@iitp.ru)[1,2],
Frolova T. I. (tfrolova@gmail.com)[1],
Iomdin L. L. (iomdin@gmail.com)[1],
Lazursky A. V. (lazursky@mail.ru) [1],
Rygaev I. P. (irygaev@gmail.com)[1],
Timoshenko S. P. (nyrestein@gmail.com)[1]

[1] A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia

[2] Universidad Politécnica de Madrid, Madrid, Spain

The paper describes a new version of the semantic analyzer SemETAP. Our approach is based on the assumption that the depth of understanding is growing with the number of inferences we can draw from the text. The salient features of SemETAP include: 1) intensive use of both linguistic and background knowledge. The former is incorporated in the Combinatorial Dictionary and the Grammar, and the latter is stored in the Ontology and Repository of Individuals. 2) Words and concepts of the ontology may be supplied with explicit decompositions for inference purposes. 3) Two levels of semantic structure are distinguished. Basic semantic structure (BSemS) interprets the text in terms of ontological elements. Enhanced semantic structure (EnSemS) extends BSemS by means of a series of inferences. 4) A new logical formalism Etalog is developed in which all inference rules are written. Semantic analysis with inference allows us to extract implicit information. The analyzer is tested on the task of interpreting high spots of the football match.

Keywords: semantic parser, ontology, inference, co-reference, question answering

СЕМАНТИЧЕСКИЙ АНАЛИЗ С ЛОГИЧЕСКИМ ВЫВОДОМ: ОСТРЫЕ МОМЕНТЫ ФУТБОЛЬНОГО МАТЧА

Богуславский И. М. (bogus@iitp.ru)[1,2],
Иомдин Л. Л. (iomdin@gmail.com)[1],
Лазурский А. В. (lazursky@mail.ru)[1],
Рыгаев И. П. (irygaev@gmail.com)[1],
Тимошенко С. П. (nyrestein@gmail.com)[1],
Фролова Т. И. (tfrolova@gmail.com)[1]

[1]Институт проблем передачи информации РАН им. А. А. Харкевича, Москва, Россия

[2]Мадридский политехнический университет, Мадрид, Испания

1. Introduction

In this paper, we describe the current state of the semantic analyzer SemETAP, different aspects of which we presented in our previous publications [Boguslavsky et al. 2015], [Boguslavsky 2017], [Rygaev 2017]. The justification of our approach and the review of the state-of-the-art were given in these publications and we will not come back to that again (except for the analysis of some recent publications in section 3).

The salient features of SemETAP are as follows.

- SemETAP is an option of the ETAP-3 linguistic processor and reuses its non-semantic modules (morphological analysis, syntactic dependency parsing, and normalization).
- Semantic analysis makes use of linguistic data and extralinguistic information (background knowledge). The linguistic data are provided by the Combinatorial Dictionary and the Grammar, and the background knowledge is stored in the Ontology and Repository of Individuals (RI). Whereas the Ontology stores hierarchically arranged information on concepts and their properties, the Repository of Individuals accumulates data on individual objects (like Moscow) or situations (like 2014 FIFA World Cup).
- Both words and concepts of the ontology may be supplied with explicit decompositions for inference purposes. We proceed from the assumption that the depth of understanding is growing with the number of inferences we can draw from the text. In many cases, a detailed description of word meanings helps produce additional inferences and thus achieve a deeper understanding.
- Semantic decomposition is carried out in terms of ontological elements. Thus, Ontology is not only a structured repository of background knowledge, but also a metalanguage for semantic description.
- Semantic analysis goes beyond the sentence boundaries. Usually, syntactic and semantic analysis of text is limited to one sentence, so that it is impossible to look

from the sentence under analysis to a neighboring one. It is however a serious obstacle for many tasks. Importantly, going beyond the sentence boundaries is essential for finding antecedents of pronouns which are very often located in one of the preceding sentences.

- Two levels of semantic structure are distinguished. Basic semantic structure (BSemS) interprets the text in terms of ontological elements. Enhanced semantic structure (EnSemS) extends BSems by means of a series of inferences.
- From the formal point of view, semantic structures of both types are represented in the RDF format, i.e. as sets of triples of the type **`relation(Ontoelement-1, Ontoelement-2)`**, where **`relation`** is an object or data property of the ontology, and **`Ontoelement-i`** is a variable or a constant denoting a concept or an instance. The RDF formalism was chosen because, on the one hand, it is very flexible and expressive, and on the other hand, it is supported by a wide range of tools and is easily integrated with many Semantic Web applications.

In previous publications, we put forward the principles which underlie the system and showed its relevant features by means of some examples. In this paper, we will give a more systematic view of the system, emphasizing the new features that were introduced lately (section 4). This section will be preceded by the problem statement (section 2) and the analysis of related work (section 3). In section 5 we will present a case study. Then we will evaluate the system (section 6) and give a brief error analysis (section 7).

2. Problem statement

Given the current state of computational semantics, semantic parsing so detailed and deep as we are aiming at is impossible to achieve for the texts of unrestricted semantics. So far, the only feasible option seems to be working with more or less narrow domains.

We believe that successive ontological, semantic and logical coverage of different domains will in the final analysis enable us to work with increasingly larger-domain texts. This approach can be illustrated by a series of studies carried out by the commonsense reasoning community, which are dedicated to the logicosemantic modeling of different domains ranging from very narrow ones (such as breaking an egg and pouring it into a bowl—[Morgenstern 2001]) to larger ones, such as emotions, interpersonal relations, commonsense psychology, causality, change of state, etc.—[Gordon, Hobbs 2004]; [Gordon, Hobbs 2011]; [Gordon, Hobbs et al. 2011]; [Hobbs, Gordon 2008]; [Hobbs, Gordon 2010]; [Hobbs, Sagae et al. 2012]; [Montazeri, Hobbs, 2011], [Montazeri, Hobbs 2012]).

In this paper, the subject domain selected is that of football reports. The texts of sports reports are an interesting objet for ontosemantic studies due to a number of features. The world of a football match is relatively small and universally understandable. This alone makes it a convenient object of modeling. On the other hand, maybe just due to the restricted size of this world, commentators and journalists do their best to make the reports less dull. This may explain high variability, a large number of individual, figurative and metaphorical expressions.

For example, the result of a match between the Russian teams Zenit and Luch-Energiya is described in one of the sites as follows:: *V matče 24 tura Čempionata Rossii po futbolu «Zenit» doma razgromil «Luč-Energiju» iz Vladivostoka so sčetom 8:1* '≈ in the match of the 24th round of the football championship of Russia Zenit defeated Luch Energiya at home with the score of 8 to 1' (Soccer.ru portal), while another site had a much more picturesque account: *Za 90 s lišnim minut igrovogo vremeni mjač pobyval v setke vorot "Luča" 8 raz, v to vremja kak vratarju "Zenita" Vjačeslavu Malafeevu prišlos' vytaskivat' sportivnyj snarjad iz setki svoix vorot liš odnaždy* '≈in slightly more than 90 minutes of playing time, the ball visited the Luch's goal net 8 times, while Zenit's goalkeeper Vyacheslav Malafeev had to pick the sporting implement out of his goal net only once' (Lenta.ru portal). The diversity of nominations used to denote the same object or situation may be striking; cf., for example, *mjač* 'ball'—*igrovoj snarjad* 'playing implement'—*sportivnyj snarjad* 'sports implement'—*sfera* 'sphere'—*kruglyj* 'the round one'. Here is how the Benfica team was referred to during one and the same report: *«Benfica»—portugal'tsy* 'the Portuguese'—*lissabontsy* 'the Lisboners'—*gosti* 'the away team'—*sopernik* 'the adversary'—*komanda* 'the team'—*portugal'skij klub* 'the Portuguese club'—*podopečnye Rui Vitoria* 'the charges of Rui Vitoria'. The goal-scoring situation is denoted in an even more diverse way. Let us give some typical examples which are far from exhausting the whole set of nominations used in the reports: *zabil* 'scored'—*zabil gol* 'scored a goal'—*otygral odin mjač* 'won one goal back'—*otkryl sčet* 'opened the scoring'—*sravnjal sčet* 'evened the score'—*vyšel vpered* 'took the lead'—*uveličil (sokratil) razryv* 'increased (reduced) the gap'—*oformil dubl'* 'made the double (scored the second goal in the match)'—*otličilsja* 'excelled'—*otpravil (poslal, zakinul, perepravil, votknul, zakatil, zapulil) mjač v setku vorot* 'sent the ball to the net'—*porazil (rasstreljal) vorota* 'hit (shot) the goal'—*realizoval penal'ti* 'realized the penalty kick'—*dobil mjač v pravyj ugol* 'dealt the final blow in the right corner'– *zamknul pas (naves)* 'closed the pass (high cross)'—*razvel mjač i golkipera po raznym uglam* 'separated the ball and the goalkeeper putting them in opposing corners'—*nakazal vratarja* 'punished the goalkeeper'—*probil mimo golkipera* 'kicked beside the goalkeeper'—*sčet stanovitsja 1:0* 'the score becomes 1:0'– *mjač okazalsja (pobyval) v setke vorot* 'the ball was in the net, visited the net'—*mjač (gol) vletel v vorota* 'the ball (the goal) flew in the goal'—*vzjatie vorot* 'seizure of the goal'—*zabil pobednyj gol* 'scored the victory goal'—*postavil pobednuju točku v matče* 'make a victory full stop in the match'—*snjal vse voprosy o pobeditele v etom matče* 'dispelled all the doubts about the winner of this match'—*emu ostavalos' tol'ko ne promaxmut'sja* 'it remained only not to miss'—*peredača (kombinatsija) okazalas' golevoj* 'the pass (combination) turned to be goal-scoring'—*zastal golkipera vrasplox* 'took the goalkeeper by surprise'—*neotrazimo probil* 'kicked irresistibly'—*ne ostavil golkiperu ni edinogo šansa* 'did not leave a single chance to the goalkeeper'—*vyjti vpered (v sčete)* 'take the lead'—*ataka zaveršilas' rezul'tativnym udarom Xondy* 'the attack ended in a successful kick by Honda'.

Many of these expressions are not synonymous. E.g. besides scoring a goal they may contain other important components of meaning. For example, *otkryt' sčet* 'open the scoring' means 'score a goal, which results in the score 1:0'. *Sravnjat' sčet* 'even the score' means 'score a goal, which results in the tie score'. *Otygrat' odin mjač* 'win

one goal back' means 'score a goal when the scoring team scored fewer goals than its adversary; as a result, this difference becomes smaller but not equal to zero'. As these examples clearly show, to adequately represent the content of many expressions semantic decomposition is an absolute must.

Besides that, it is characteristic of sports reports to recur to the indirect mode of expression. Many meaning components are expressed implicitly, and the text interpretation system should be able to restore them. Let us give a typical example.

(1) *Korner u vorot xozjaev polja zaveršaetsja udarom Netsida v upor, no Dikan› okazyvaetsja na vysote* 'the corner kick at the goal of the home team ended in the kick point blank by Necid, but Dikan was up to the mark'.

If the Hearer is aware of the background information, he will easily understand that Necid failed to score a goal, although this was not said directly. We will come back to this example below (in 4.3) and show how SemETAP manages to cope with it.

All of the aforesaid makes sports reports understanding a linguistically nontrivial and exciting task. In processing football reports, we lay emphasis on the understanding of "high spots" of the match, similar to (1). We call high spots the moments fraught with scoring a goal, for example when the goal of one of the teams is being attacked. Our aim is to identify major details of the situation making use of all the information available.

3. Related work

Although football is a popular topic of computational linguistics experiments, most of the relevant efforts have been focused on a football ontology construction (cf. Tsinaraki et al. 2005, Schmidt 2006, Abreu et al. 2010, Ranwez Soccer Ontology, SWAN Soccer Ontology) or on generating football match summaries using an ontology (cf. Nadjet Bouayad-Agha et al. 2011). A notable exception is a recently published book Cimiano et al. 2014, which is also using football as its subject domain. At the level of foundational principles, the approach defended in this book is very similar to ours. It proclaims that in order to interpret natural language texts with respect to the domain knowledge, a machine needs (a) a formalization of the domain knowledge by means of an ontology, (b) a process for building meaning representations that are aligned to that domain knowledge, and (c) a way to draw inferences and use the resulting information in the interpretation process. We cannot agree more with these theses. However, their implementation in Cimiano et al. 2014 and in our project is quite different. Besides, it remained unclear to us up to what extent these principles have been implemented in a real system. In particular, it was difficult to make an idea of syntactic and semantic complexity of sentences the system copes with.

One of the differences between our approach and the one of Cimiano and his co-authors is that their ontology does not support the representation of events and their modification (Cimiano et al. 2014 48–49). As we will show below, our language of meaning representation is much more expressive.

Another important difference between our approaches concerns the role played by the ontology and the status of the NL dictionary. The Cimiano approach is radically

ontology-centric. According to it, each text belongs to some specific domain, and one should first of all create an ontology of this domain and then compile a NL dictionary whose role is to specify NL equivalents for the ontological elements. A domain-independent dictionary also exists but its scope is limited to representing closed-class words such as determiners, pronouns, auxiliary verbs, etc. Dictionaries induced by different domain ontologies will be different even in the number and granularity of meanings of particular words.

We cannot accept this approach. It implies that there is no such a thing as a dictionary of a particular language. There are as many dictionaries as specific domains (such as the football domain), which may contain the same words that have different meaning sets and different granularity. We think that such an approach will be very difficult to implement, since there is no clear-cut border between the vocabularies of different domains, as well as between domain-specific and general vocabulary. Besides, it is often very difficult to assign a text to a specific domain. Then which dictionary should be used for its processing? In our opinion, the domain-independent dictionary should not be restricted to closed-class words, since even domain-specific texts contain a large number of general vocabulary words.

We adopted a different approach. We have an integrated dictionary for Russian, in which all domain-specific information is marked in a special way. Such a marking is needed not only for the words that do not occur beyond domain-specific texts. Very often, it is only some senses of a word (or some phrases containing this word) that are domain-specific, other senses being quite neutral. For example, the phrase *red card* can be easily encountered in a free text where it merely means a card whose color is red. Yet in football it denotes a specific punishment and corresponds to a concept (**`RedCard`**) of the ontology. The connection between the phrase *red card* and this concept is marked as relevant for the sports domain (DOMAIN:SPORT-DOMAIN). As an illustration, below is a fragment of the dictionary entry for КАРТОЧКА ‘card’.

ENTRY:КАРТОЧКА
…
    ZONE:EN
        TRANS: CARD
    ZONE:SEM
        DOMAIN:SPORT-DOMAIN
<a rule stating that *red card* corresponds to **`RedCard`**>
…

If the text we are processing belongs to this domain, the **`RedCard`** interpretation will be preferred. Otherwise, it will have the status of only one of possible alternatives. This strategy allows us to have a single dictionary matched with one or more domain-specific ontologies.

4. Semantic analyzer SemETAP

In its present state, the SemETAP analyzer is a follow-up of the system described in Boguslavsky et al. 2015, Boguslavsky 2017 and Rygaev 2017. Below, we will show what it looks like today with a particular focus on the components developed recently.

4.1. Analysis of football reports

At the input, SemETAP receives the Normalized Syntactic Structure (NormSynS), constructed by the regular ETAP-3 parser. By this moment, all strongly governed prepositions and conjunctions, as well as auxiliary verbs have been deleted, zero copulas have been substituted by the verb *byt'* 'to be', lexical functions (such as Oper, Func and others) have been identified, antecedents of anaphoric pronouns have been found and some other normalization operations have been performed. Further, NormSyntS is subjected to three stages of processing: 1) preparation of NormSyntS for semantization, 2) construction of BSemS, 3) construction of EnSemS.

4.1.1. Preparation of the Normalized Syntactic Structure for semantization

At this stage, the following operations are carried out, among others:

- Substitution of antecedents for anaphoric pronouns and making explicit zero actants.

  Pust' vratar' sygral i ne očen' uverenno, no ugrozu ot svoix vorot on [⇒ *vratar'*] *otvel.* '≈ Even though the goalkeeper did not play very strongly but he [⇒ the goalkeeper] fended of the threat to his goal'

  Traore skinul mjač pod udar Ionovu, kotorogo [⇒ *Ionova*] *v poslednij moment operedil Samba.* '≈ Traore kicked the ball to Ionov who [⇒ Ionov] was outrun by Samba at the last moment'

- Resolving non-anaphoric coreference based on the background knowledge extracted from the Repository of Individuals.

  Dumbia, obygrav neskol'kix sopernikov, vyvel Tošiča odin na odin s Fil'tsovym, posle čego serbu [⇒*Tošiču*] *ostavalos' tol'ko ne promaxnut'sja.* . '≈ Dumbia who outplayed several adversaries brought Tošič head to head with Filtsov, after which the Serb [⇒*Tošič*] only needed not to miss'

- Processing of support verbs aiming at obtaining identical BSemSs for sentences like:

  Spartak pobedil Dinamo 'Spartak defeated Dinamo' = *Spartak oderžal pobedu nad Dinamo* 'Spartak gained a victory over Dinamo' = *Spartak nanjos poraženie Dinamo* 'Spartak inflicted a defeat to Dinamo' = *Dinamo poterpelo poraženie ot Spartaka* 'Dinamo suffered a defeat from Spartak'.

- Splitting the sentence into predications (subordinate, participial, infinitival clauses, predicative NPs).

  Sentence *V seredine pervogo tajma Netsid posle pasa Dumbija bjet po vorotam, i Malafeev s trudom perevodit mjač na uglovoj, posle podači kotorogo ivuariets*

popadaet v perekladinu '≈ In the middle of the first period, Netsid, after a pass by Doumbia kicks the ball towards the goal, Malafeev, with difficulty, moves the ball over the goal line to enable a corner kick, so that, after the corner was kicked, the Ivorian hits the crossbar' is represented by means of 5 temporally ordered predications: 1) Doumbia gives a pass, 2) Necid kicks the ball towards the goal, 3) Malafeev moves the ball over the goal line, which results in a corner kick, 4) somebody kicks the corner, 5) the Ivorian hits the crossbar.

- Transformation of the passive voice into the active one.

  Rossijskij futbol byl predstavlen v plej-off srazu dvumja komandami 'Russian football was represented at the play-off by two teams at once' ⇒ *Srazu dve komandy predstavljali v plej-off rossijskij futbol* 'two teams at once represented Russian football at the play-off'.

4.1.2. Constructing Basic Semantic Structure

Basically, this stage contains semantic interpretation of words, syntactic constructions and morphological features by means of ontological elements. If a word has an exact equivalent among the ontology concepts, it is replaced with this concept. If needed, this concept will be semantically interpreted at the next stage. For example, *gol* 'goal'

⇒ **`GoalEvent`**.

If the ontology does not have such an equivalent, and it is inexpedient to create it, then a rule is composed which constructs a fragment of BSemS. For example, *vratar'* 'goalkeeper' is translated as **`Human hasRole GoalkeeperRole`** ("person that fulfills the goalkeeper role").

A more complicated rule is responsible for interpreting relational adjectives such as *frantsuzskij* 'French'. For readers' convenience, we will not reproduce it here in the formal language, but give its simplified NL gloss:

1) If *frantsuzskij* 'French' modifies a noun which corresponds to the ontological class **`SportAgent`**, then *frantsuzskij* translates as 'representing France' (e.g. *frantsuzskaja sbornaja* 'French national team').
2) If *frantsuzskij* 'French' modifies a noun which corresponds to the ontological class **`Human`**, then *frantsuzskij* translates as 'living in France' (e.g. *frantsuzskie bolel'ščiki* 'French fans').
3) If *frantsuzskij* 'French' modifies a noun which corresponds to the ontological class **`Organization`**, then *frantsuzskij* translates as 'acting in France' (e.g. *frantsuzskij muzej* 'French museum').
4) If *frantsuzskij* 'French' modifies a noun which corresponds to one of the ontological classes **`OrganizedEvent`**, **`Building`**, **`StationaryArtifact`**, **`GeographicArea`**, then *frantsuzskij* translates as 'situated in France' (e.g. *frantsuzskie bul'vary (reki)* 'French boulevards (rivers) ').
5) If *frantsuzskij* 'French' modifies a noun which corresponds to one of the ontological classes **`Document`**, **`Food`**, **`Artifact`**, then *frantsuzskij* translates as 'made in France' (e.g. *frantsuzskoe vino* 'French wine').

4.1.3. Constructing Enhanced Semantic Structure

The rules that operate at this stage mostly explicate the semantics of concepts. To give an example, it is not sufficient to state that all numerous ways of denoting goal scoring correspond to the concept **`GoalEvent`** (this task is already solved in BSemS). It is no less important to show what exactly goal scoring is. Briefly, a goal is scored if a player of team A kicks the ball with the aim of moving it in the goal of team B; as a result, the ball gets into the goal of team B and the score of team A increases by 1; this event is beneficial for team A and unbeneficial for team B.

Obviously, such a decomposition enables us to obtain a much deeper comprehension of the text and to make more inferences, than if we restricted ourselves to merely establishing the fact that **`GoalEvent`** takes place. For instance, we can infer what the player of team A has done, where the ball is located just after the event, what happened to the score, who benefitted from the event, etc. By the same token, we can better understand texts like *Udar pjatkoj, i mjač v setke vorot* 'a kick with the heel, and the ball is inside the goal'. We are informed that a goal has been scored although *goal scoring* has not been mentioned.

In describing events, special attention is paid to such aspects as preconditions of the event and its results, both obligatory and possible, objectives of the participants, the actions they perform, the assessment of the event from the point of view of different participants, etc.

It is to be stressed that predications may have different degrees of epistemic modality. In particular, the maximal degree (**`EpistModality hasDegree MaximalDegree`**) is assigned to an event that definitely took place. The medium degree (**`EpistModality hasDegree MediumDegree`**) corresponds to a possible event. Due to this, we can differentiate between the 100%-reliable logical entailments and the inferences that are no more than plausible expectations. The importance of the latter for the interpretation of discourse and, in particular, dialogues, is exemplified in [Boguslavsky et al. 2016].

At the time of writing (February, 2018), we dispose of 261 rules for transforming BSemS into EnSemS. Some of these rules are related to the general vocabulary, and others describe domain (football) concepts. Even the general vocabulary is far from being completely covered by the rules, to say nothing of the concepts of other domains, so that the inventory of rules should be significantly augmented in the future.

These rules are written in a special language, which will be discussed in the next section.

4.2. Language for inference and inference rules

When we were selecting a formalism for writing inference rules, we came to the conclusion that none of the existent formalisms we were aware of could be directly used for our task. We decided that it would be better to develop a formalism of our own, which would be sufficiently expressive for defining the meaning of the concepts and at the same time allow for efficient implementation of the algorithms for logical inference. In developing such a formalism, the following requirements were taken into account:

1. The formalism is destined for formulating logical inference rules. An important particular case of such rules are rules for decomposing the meaning of concepts.
2. The rules should apply to the semantic structure represented by an RDF graph.
3. The rule application should result in adding new triples to SemS. They embody new knowledge obtained by logical inference.
4. The rule application should be efficient from the point of view of the system productivity.
5. The formalism should be easily understandable for the linguists.

As a result of taking these requirements into account, the Etalog language was born, which combines some elements of Datalog (requirements 1 and 4), RDF/SPARQL (requirements 2 and 3) and natural language (requirement 5).

A rule in Etalog consists of a title (which contains the keyword Rule and the name of the rule), a logical premise, the implication sign (->) and a conclusion. The premise and the conclusion are one or more predicates separated by a comma (stands for conjunction). The rule terminates with a full stop.

Etalog is used to write BSemS ⇒ EnSemS rules. Here is an example of one of these rules. It interprets the concept **`EqualToOccasion`** and is used for processing sentence (1), which we already mentioned in section 2 and will discuss in more detail below (symbol // introduces a NL comment):

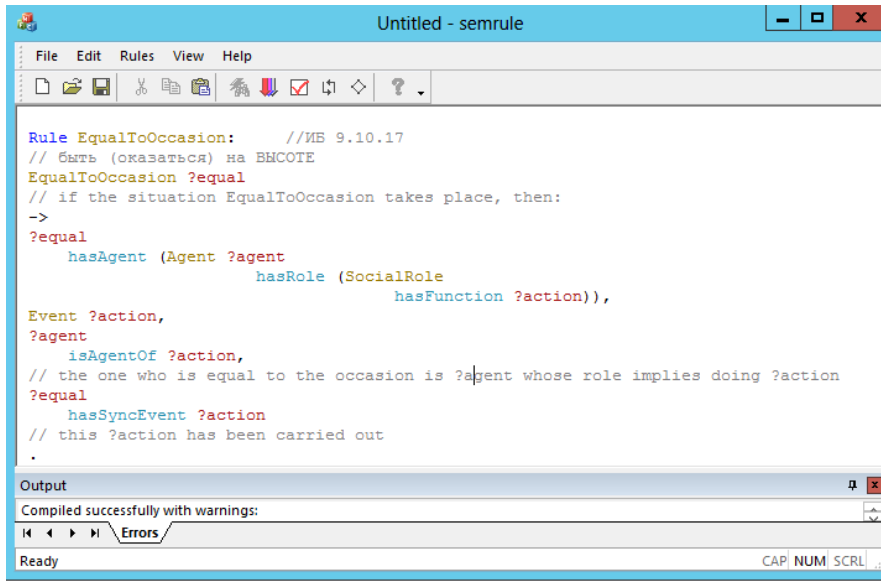


Fig. 1. A semantic rule written in Etalog

4.3. Reasoner

Rules written in Etalog are applied by the RDFox reasoner developed at Oxford University (Motik 2014, Nenov 2015). We chose this reasoner because it suits our needs very well in a number of ways:

1. Very efficient and scalable reasoner. Shows the top results in benchmarking even for large datasets (Benedikt 2017).
2. Optimized for RDF model. Provides an RDF triple store and supports efficient SPARQL query execution.
3. Supports new variables in the consequent of the rules not present in the antecedent. Such variables are known as existentials or anonymous individuals and are essentially required in concept definition rules to represent various parts of the definition.
4. Supports controlled materialization of the existentials (adding new individuals and relations to the semantic structure) allowing for custom filters to prevent infinite loops and guarantee termination. Such procedure is known in the literature as the 'restricted chase' (Benedikt 2017, p. 40).
5. Has a built-in support for equality relation (Motik 2015) and a special query mode where different but equal individuals are treated as one individual. It is very helpful in coreference processing.
6. Originally written in C++ and has a solution for Windows. So it integrates very well with ETAP which is also written in C++ for Windows.

Each Etalog rule is translated into several RDFox inference rules. The rule is split into several chunks which are applied independently. First of all functional relations are extracted from the Etalog rule consequent and a separate RDFox rule is created for each of them. The rest of the consequent is split into independent chunks and an RDFox rule is created for each chunk.

This is done to maintain integrity and avoid creation of duplicated entities. Each RDFox rule includes a filter—it does nothing if the corresponding subgraph already exists. The filter works at the level of RDFox rules, so for an Etalog rule it is possible that some individuals will be accommodated from the existing data while others will be added (see more details in Rygaev 2017). This happens invisibly for the linguists who create rules in Etalog, so they can concentrate on the concept definition and can ignore technical aspects of the rule application.

The filters do not always prevent creation of duplicated objects. Because of that we also use equality rules to join duplicated objects together. First of all such rules are created automatically for each functional relation. For more complex cases, additional equality rules can be written manually in Etalog.

We also have an additional filter to guarantee termination of the reasoning. If all the variables in the antecedent of the RDFox rule are anonymous (i.e. do not come from the original data but are created by other rules) the rule does nothing. This is required to avoid infinite chains such as the following: if a player controls the ball he can pass it to another player who then will be controlling the ball and will be able to pass it to another player and so on ad infinitum. This empirical filter works surprisingly well, preventing unnecessary inferences and very rarely blocking good inferences.

4.4. Repository of Individuals

We have built a large Repository of Individuals, which contains data on more than 200K individuals automatically extracted from DBpedia. These individuals belong to the following ontology classes: **`Human`**, **`FootballTeam`**, **`TimeInterval`**, **`IndependentState`**, **`City`**, **`SportsLeague`**. The data on the football players include the name, family name, place and date of birth, country of residence, team (or teams) he played for during his carrier, playing position in the team, etc.

Let us show how all the three resources—Combinatorial Dictionary, Ontology and the Repository of Individuals—contribute to the interpretation process. Let us go back to sentence (1) referred to at section 2.

(1) *Korner u vorot xozjaev polja zaveršaetsja udarom Netsida v upor, no Dikan' okazyvaetsja na vysote* 'the corner kick at the goal of the home team resulted in the kick point blank by Necid, but Dikan was up to the mark'.

We want to know if a goal has been scored. To answer this question, we will have to recur to three sources of information:

- Combinatorial Dictionary tells us that the expression *byt' na vysote* 'be up to the mark' corresponds to the concept **`EqualToOccasion`**, interpreted as 'do well what one is expected to do' (cf. the rule in the previous section);
- Repository of Individuals contains the information that Andrei Dikan is a goalkeeper of Spartak Football Club;
- Ontology describes the goalkeeper role as preventing the ball from penetrating the goal of his team.

These three pieces of information allow the reasoner to infer that Dikan, being a goalkeeper, performed well his function of preventing a goal and, consequently, a goal has not been scored. Obviously, if the Repository of Individuals had told us that Dikan had the position of a forward, then, given that the Ontology specifies the function of a forward as scoring goals, the overall conclusion would have been opposite.

Again, a conclusion concerning scoring a goal has been made in the context which does not mention the word *goal* nor any of its synonyms.

5. Case study

Let us give another example to illustrate the interpretation of a sentence by means of a series of inferences. We will analyze sentence (2) and show how the analyzer comes to the conclusion that the team for which Aršavin was playing has suffered a defeat.

(2) *Aršavin tak i ne smog spasti matč* 'Aršavin could not save the match'.

Among the data at the disposal of the analyzer there are the following three facts which we will for the readers› convenience formulate in NL and not in Etalog, in which they are stored in the system:

1. The verb *smoč* 'be able', in the perfective aspect, is implicative (for more details on the implicative verbs in Russian and the impact the verbal aspect

has on the implicativity cf. Rygaev 2015). Therefore, *X smog P* 'X could do P' implies that P took place, while *X ne smog P* 'X could not do P' implies that P did not take place.

2. The phrase *spasti matč* 'save the match' is interpreted as 'prevent the defeat of one's team'.
3. 'Prevent' is also an implicative predicate, but of a different type than 'be able'. *X prevented P* implies that P did not take place.

These facts underlie the following inference chain: Aršavin could not save the match ⇒ does not take place: Aršavin saved the match ⇒ does not take place: Aršavin prevented the defeat of his team ⇒ does not take place: the team for which Aršavin played was not defeated ⇒ the team for which Aršavin played was defeated.

Let us take a look how the system formally makes these inferences.

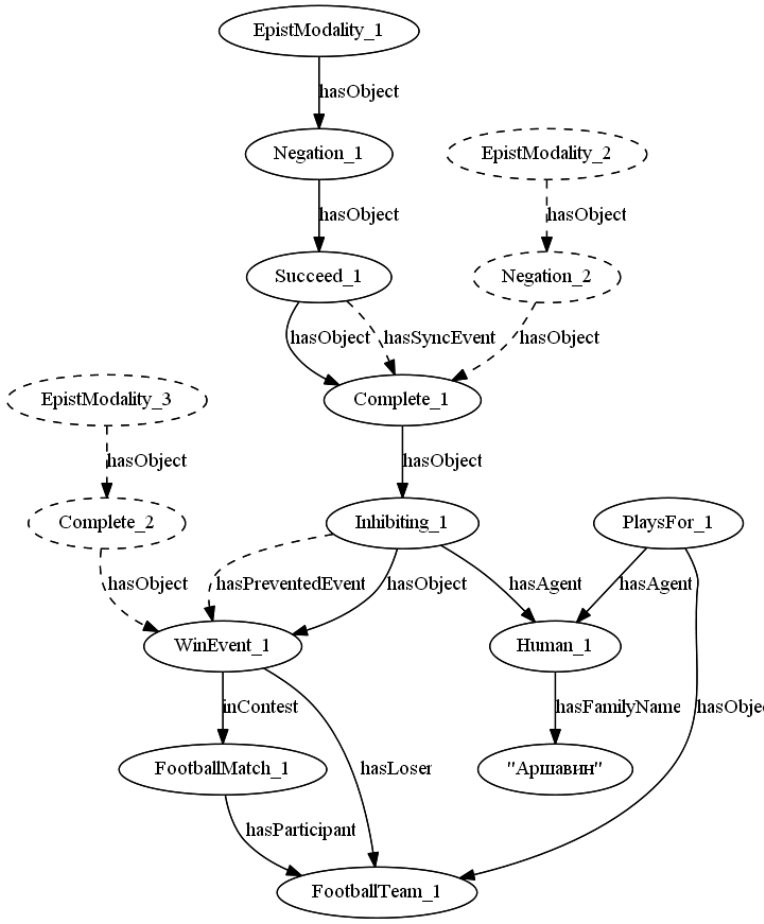


Fig. 2. Elements of BSemS (solid) and EnSemS (dashed) of the sentence *Aršavin tak i ne smog spasti matč* (Aršavin could not save the match)

In Fig. 2 we can see in solid lines the relevant elements of the BSemS which are created based on lexical and grammatical properties of the sentence. Nodes **Negation _ 1 hasObject Succeed _ 1** correspond to the lexical items *ne smog* 'could not'. Nodes **Human _ 1 hasFamilyName "Аршавин"** correspond to *Aršavin*. **Complete _ 1** comes from the perfective aspect of *spasti* 'save'. **EpistModality _ 1** marks the top predicate of the sentence indicating its facticity status. The rest (**Inhibiting _ 1** and down) comes from the lexical meaning of *spasti matč* 'save the match' and is created by a dictionary rule of the verb SPASAT 'save' shown below:

```
ZONE:SEM
  DOMAIN:SPORT-DOMAIN
    REG:SEM-CONV2.D0
      TAKE:*
      CHECK
        1.1 DOM-EQUN(X,Z,ПРЕДИК,Human)
      N:1 // Иванов спас матч
        CHECK
          1.1 DOM-LEXR(X,W,1-КОМПЛ,МАТЧ)
        DO
          1 REPLACE-SEM(X,Inhibiting)
          2 REPLACE-SEM(W,FootballMatch)
          3 ADD-NODE-SEM(Z1,WinEvent)
          4 REPLACE-LINK([X,Z,*],[X,Z,hasAgent])
          5 LINK-NODES(X,Z1,hasObject)
          6 ADD-NODE-SEM(U,FootballTeam)
          7 REATTACH-NODE([X,W,*],[Z1,W,inContest])
          8 LINK-NODES(Z1,U,hasLoser)
          9 ADD-NODE-SEM(U3,PlaysFor)
          10 LINK-NODES(U3,Z,hasAgent)
          11 LINK-NODES(U3,U,hasObject)
          12 LINK-NODES(W,U,hasParticipant)
```

Once BSemS is created, inference rules are applied. First the definitions of the concepts **Succeed** and **Inhibiting** are processed, adding two relations **hasSyncEvent** and **hasPreventedEvent**. **hasSyncEvent** means that two events have the same facticity status, while **hasPreventedEvent** means that they have the opposite facticity status. The corresponding parts of the definitions are presented below:

```
Succeed ?x ->
?x      hasObject (Event ?event)
        hasSyncEvent ?event.
Inhibiting ?x ->
?x      hasObject (Event ?event)
h       asPreventedEvent ?event.
```

Then the inference rule for **hasSyncEvent** relation creates **EpistModality _ 2** and **Negation _ 2** nodes thus marking the fact that **Inhibiting _ 1** did not take place based on the fact that **Succeed _ 1** did not take place.

```
?event1 hasSyncEvent ?event2,
      EpistModality hasObject (Negation hasObject ?event1) ->
      EpistModality hasObject (Negation hasObject ?event2).
```

And finally the inference rule for **hasPreventedEvent** relation creates **EpistModality _ 3** and **Complete _ 2** nodes thus marking the fact that **WinEvent _ 1** took place based on the fact that **Inhibiting _ 1** did not take place.

```
?event1 hasPreventedEvent ?event2,
      EpistModality hasObject (Negation hasObject(Complete
      hasObject ?event1)) ->
      EpistModality hasObject (Complete hasObject ?event2).
```

Since EnSemS contains a very large number of predications (up to several hundred) and is difficult to survey, the most convenient way to make sure that the analyzer obtained the expected inference is the question-answering option. In this option, the analyzer constructs the EnSemS of both the initial sentence, and the question, transforms the EnSemS of the question into SPARQL and infers the answer with the help of the RDFox reasoner. In Fig. 3 one can see the result of processing sentence (2) in the question-answering mode. In the upper window is the text (*Aršavin could not save the match*) and the diagnostic question (*Did Aršavin's team lose the match?*). The lower window contains EnSemSs of both sentences (only the last lines of the EnSemS of the question are seen) and the answer returned.

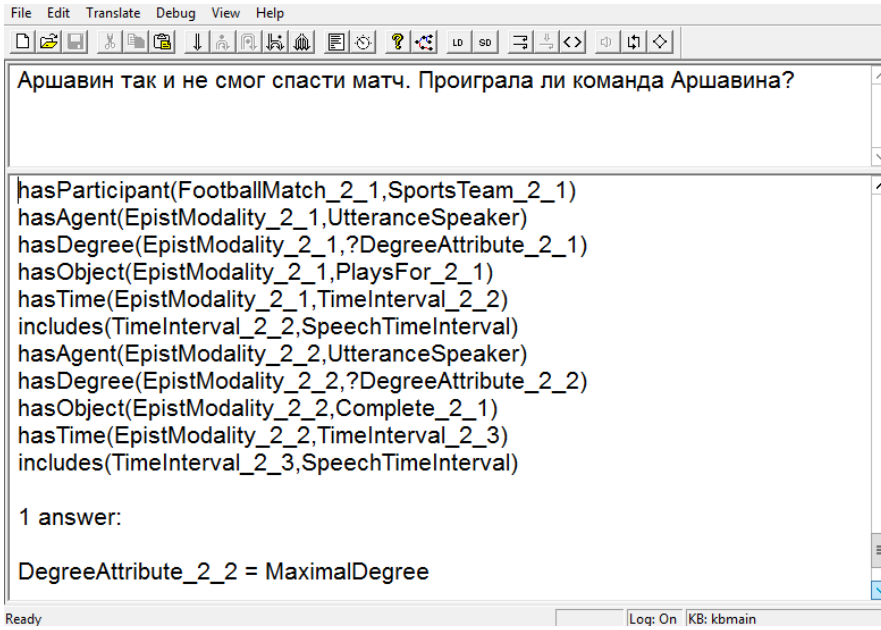


Fig. 3. EnSemS of sentence (2) and of the question and the answer obtained

EnSemS of the question can be glossed as follows: what is the value of the epistemic modality of the statement "the team for which Aršavin was playing was defeated"? In plain words, it means: is it true that Aršavin's team was defeated? The answer, which can be seen in the lower window, reads that the value of this modality is maximal. This means that the question was answered in the affirmative.

6. Evaluation

For the evaluation, we selected 50 sentences the analyzer did not see before which are having to do with high spots of the match, i.e. the moments fraught with scoring a goal, such as attacks on the goal, direct free kicks, corners, goalkeeper interventions, etc. The sentences were extracted from the online running commentaries on the Football World Cup qualifying games. The analyzer constructed full EnSemS of all the sentences. The aim of the evaluation was to determine up to what extent the analyzer was able to extract explicit or infer implicit information on the main features of the high point. Concretely, we found that:

- 11 out of 50 test file sentences describe the situations that ended in scoring a goal. In 10 cases (91%) the semantic analyzer successfully identified the goal event, in 5 cases (45.5%) it also was able to identify the author of the goal event.
- 7 sentences out of 50 specify the distance from which a goal scoring shot was performed, and in 6 cases (85.7%) the semantic analyzer showed the correct distance.
- 12 sentences of the test file contain the information about the so-called "starting point" of the shot (which corresponds to the location of the person performing the shot). In 11 cases (91.6%) the EnSemS indicate the starting point correctly.
- The terminal point of the shot (the part of the goal the shot is aimed at) is mentioned in 17 sentences. In 16 cases (94%) EnSemS indicate it correctly.
- Whenever the goalkeeper managed to prevent the goal event (6 sentences out of 50), the semantic analyzer showed the correct result (100%).

7. Error analysis

The quality of semantic structures strongly depends on the accuracy of syntactic structures they are built on. 9 syntactic structures obtained for the test file contained some syntactic errors. Since our aim was to evaluate the semantic component of the system, we performed some interventions to the syntactic component in order to correct these errors.

The defects encountered in EnSemS are of the following types.

1. Wrong generation of an explicit subject in a noun phrase with a zero subject.
2. Wrong interpretation of the NP *štrafnaja X-a* 'X's penalty box'. The rule assumes that such a NP is only appropriate if X is the goalkeeper of the team to whom the penalty box belongs. However, one of the sentences of the test file refutes this supposition: in the sentence *"Ferreira Carrasco made a fault near his penalty box"* Ferreira Carrasco is not a goalkeeper.

3. The resolution of the non-anaphoric co-reference (*US President—Donald Trump*) requires that the individual in question be represented in the Repository of Individuals. Some of the players referred to in the test file are absent from RI, and therefore the co-reference between their mentions was not established.
4. Verbal tense is not always interpreted correctly. In Russian, the present tense may have a so called "commentary interpretation" (*nastojaščee reportažnoe*). The sentence *Ečše odin mjač zabivajut bel'gijtsy* 'Belgium scores one more point' denotes a terminated event although its verb is in the present imperfective. This use of present is typical of the genre of commentaries. Our rules fail to distinguish between the regular present and the present of commentaries.

Conclusions

The SemETAP semantic analyzer is an option of the ETAP-3 Linguistic Processor aiming at producing in-depth semantic interpretation of the Russian text. SemETAP makes use of both linguistic and extra-linguistic (background) knowledge, the former being stored in the Combinatorial Dictionary and the Grammar, and the latter—in the Ontology and the Repository of Individuals. Semantic analysis represents the text on two levels: Basic semantic structure (BSemS) interprets the text in terms of ontological elements. Enhanced semantic structure (EnSemS) extends BSemS by means of a series of inferences. An important feature of the analyzer is its capacity to infer implicit information, which is very useful for a variety of applications including question answering, story understanding, and dialogue processing.

Acknowledgements

This work was supported by the RSF grant 16-18-10422, which is gratefully acknowledged.

References

1. *Abreu, P. Faria, M., Reis L., Garganta, J.* (2010), "Knowledge Representation in Soccer Domain: An Ontology Development". 5th Iberian Conference on Information Systems and Technologies (CISTI), 2010.
2. *Benedikt, M., Konstantinidis, G., Mecca, G., Motik, B., Papotti, P., Santoro, D., and Tsamoura, E.* (2017). Benchmarking the chase. In Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (pp. 37–52). ACM.
3. *Boguslavsky I.* (2017), Semantic Descriptions for a Text Understanding System. In: Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2017), p. 14–28.
4. *Boguslavsky I., V. Dikonov, Frolova T., L. Iomdin, A. Lazursky, I. Rygaev, V. Sizov, S. Timoshenko.* (2016), Plausible Expectations-Based Inference for Semantic Analysis // Proceedings of the 2016 International Conference on Artificial Intelligence (ICAI"2016). USA: CSREA Press, 2016. pp. 477–483.

5. *Boguslavsky I., V. Dikonov, L. Iomdin, A. Lazursky, V. Sizov, S. Timoshenko.* (2015), Semantic Analysis and Question Answering: a System Under Development. In: Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2015), p.62.
6. *Bouayad-Agha, N., Casamayor, G., Wanner, L., Díez, F., López Hernández, L.* (2011), "FootbOWL: Using a generic ontology of football competition for planning match summaries". Extended Semantic Web Conference ESWC 2011: The Semantic Web: Research and Applications, pp 230–244.
7. *Cimiano Ph., Unger Ch., McCrae J.* (2014), Ontology-based Interpretation of Natural Language. Synthesis Lectures on Human Language Technologies. Morgan and Claypool Publishers.
8. *Gordon, Andrew S., and Jerry R. Hobbs.* (2004), "Formalizations of Commonsense Psychology", AI Magazine, Winter 2004, pp. 49–62.
9. *Gordon, Andrew S., and Jerry R. Hobbs.* (2011), "A Commonsense Theory of Mind-Body Interaction", in Proceedings of the 10th Symposium on Logical Formalizations of Commonsense Reasoning, AAAI Spring Symposium Series.
10. *Gordon, Andrew S., Jerry R. Hobbs, and Michael T. Cox.* (2011), "Anthropomorphic Self-Models for Metareasoning Agents", in Michael T. Cox and Anita Raja (eds.), Metareasoning: Thinking about Thinking, The MIT Press, Cambridge, Massachusetts, pp. 295–305.
11. *Hobbs, Jerry R., and Andrew Gordon.* (2008), "The Deep Lexical Semantics of Emotions", Proceedings, LREC 2008 Workshop on Sentiment Analysis: Emotion, Metaphor, Ontology, and Terminology, Marrakech, Morocco, May 2008.
12. *Hobbs, Jerry R., and Andrew Gordon.* (2010), "Goals in a Formal Theory of Commonsense Psychology", in A. Galton and R. Mizoguchi (eds.), Formal Ontology in Information Systems: Proceedings of the Sixth International Conference (FOIS 2010), IOS Press, Amsterdam, pp. 59–72.
13. *Hobbs, Jerry R., Alicia Sagae, and Suzanne Wertheim.* (2012), "Toward a Commonsense Theory of Microsociology: Interpersonal Relationships", in M. Donnelly and G. Guizzardi (eds.), Formal Ontology in Information Systems: Proceedings of the Seventh International Conference (FOIS 2012), IOS Press, Amsterdam, Netherlands, pp. 249–262.
14. *Montazeri, Niloofar, and Jerry R. Hobbs.* (2011), "Elaborating a Knowledge Base for Deep Lexical Semantics", in J. Bos and S. Pulman (eds.), Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011), January 2011, pp. 195–204.
15. *Montazeri, Niloofar, and Jerry R. Hobbs.* (2012), "Axiomatizing Change-of-State Words", in M. Donnelly and G. Guizzardi (eds.), Formal Ontology in Information Systems: Proceedings of the Seventh International Conference (FOIS 2012), IOS Press, Amsterdam, Netherlands, pp. 221–234.
16. *Motik, B., Nenov, Y., Piro, R., Horrocks, I., and Olteanu, D.* (2014). Parallel Materialisation of Datalog Programs in Centralised, Main-Memory RDF Systems. In AAAI (pp. 129–137).
17. *Motik, B., Nenov, Y., Piro, R. E. F., and Horrocks, I.* (2015). Handling Owl: sameAs via Rewriting. In AAAI (pp. 231–237).

18. *Morgenstern, Leora.* (2001), MidSized Axiomatizations of Commonsense Problems: A Case Study in Egg Cracking. Studia Logica, April 2001, Volume 67, Issue 3, pp 333–384
19. *Nenov, Y., Piro, R., Motik, B., Horrocks, I., Wu, Z., and Banerjee, J.* (2015). RDFox: A highly-scalable RDF store. In International Semantic Web Conference (pp. 3–20). Springer, Cham.
20. *Rygaev I.* (2015), Implicative verbs in Russian and their semantic analysis in ETAP-3 linguistic processor. [Implikativnye glagoly v russkom jazyke i ix semantičeskij analiz v ramkax lingvističeskogo protsessora ETAP-3]. Master thesis, RGGU, 2015.
21. *Rygaev I.* (2017), Rule-based Reasoning in Semantic Text Analysis. Proceedings of the Doctoral Consortium, Challenge, Industry Track, Tutorials and Posters @ RuleML+RR 2017 hosted by International Joint Conference on Rules and Reasoning 2017 (RuleML+RR 2017).
22. *Schmidt, Thomas.* (2006), "Interfacing Lexical and Ontological Information in a Multilingual Soccer FrameNet". Proceedings of OntoLex 2006—Interfacing Ontologies and Lexical Resources for Semantic Web Technologies.
23. *Tsinaraki, C., Polydoros, Kazasis, F., and Christodoulakis, S.* (2005), Ontology-based semantic indexing for mpeg-7 and tv-anytime audiovisual content. Multimedia Tools and Applications, Vol. 26, Num .3, 2005, pp. 299–325.

TERM EXTRACTION FOR CONSTRUCTING SUBJECT INDEX OF EDUCATIONAL SCIENTIFIC TEXT

Bolshakova E. I. (eibolshakova@gmail.com)
Moscow State Lomonosov University, National Research University Higher School of Economics, Moscow, Russia

Ivanov K. M. (ivanov.kir.m@yandex.ru)
Moscow State Lomonosov University

Subject index, or back-of-the-book index, is a device intended to provide an easy access to relevant fragments of a text document. Subject indexes usually contain particular single-word and multi-word terms from the corresponding documents. Such indexes are especially useful for reading large documents with specialized terminology, as well as educational texts in difficult scientific and technical areas. The central problem of back-of-the-book indexing is recognition of terms to be included into the index. The paper describes a method developed for extracting and filtering terms from a given educational scientific text, with the purpose of reliable term selection in computer indexing systems. The method is primarily based on rules with lexico-syntactic patterns representing linguistic information about terms and typical contexts of their usage in Russian scientific and educational texts; simple occurrences statistics of terms is used as well. Experimental evaluation of the method has shown a considerable increase of precision and recall of term extraction compared with the widely-used standard techniques.

Keywords: rule-based term extraction, back-of-the-book index, subject indexing, educational scientific texts, lexico-syntactic patterns

ИЗВЛЕЧЕНИЕ ТЕРМИНОВ ДЛЯ ПОСТРОЕНИЯ ПРЕДМЕТНОГО УКАЗАТЕЛЯ УЧЕБНО-НАУЧНОГО ТЕКСТА

Большакова Е. И. (eibolshakova@gmail.com)
МГУ имени М. В. Ломоносова, НИУ ВШЭ, Москва, Россия

Иванов К. М. (ivanov.kir.m@yandex.ru)
МГУ имени М. В. Ломоносова, Москва, Россия

Предметный указатель к текстовому документу обычно содержит значимые однословные и многословные термины текста, вместе с номерами страниц, где они встречаются, что облегчает доступ к нужным фрагментам документа. Предметные указатели особо полезны для больших документов со специальной терминологией, а также для учебных текстов в сложных научных и технических областях. Центральной проблемой построения предметного указателя является выявление и отбор терминов для включения в указатель. В статье описывается метод, разработанный для извлечения из текста терминов и их последующей фильтрации в рамках программной системы поддержки построения предметных указателей. Метод основан на применении правил с лексико-синтаксическими шаблонами, отражающими лингвистическую информацию о терминах и контекстах их использования в учебно-научных текстах на русском языке. Экспериментальная оценка метода показала существенный прирост точности и полноты отбора терминов по сравнению с широко используемой стандартной технологией извлечения терминов из текстов.

Ключевые слова: извлечение терминов на основе правил, предметный указатель, учебно-научные тексты, лексико-синтаксические шаблоны

1. Introduction

Subject, or back-of-the-book, indexes are often constructed for large and medium-size text documents, such as books, manuals, tutorials etc., especially in highly specialized domains. As a rule, subject indexes contain significant terms from the corresponding documents, with associated page numbers. Such indexes are usually placed at the back of the text documents in order to facilitate navigating through them and locating needed information. Typical fragments of subject indexes are presented in Figure1 (hierarchical index in English and flat index in Russian).

… …
– B –
binary file 67
bit 7
block diagram 20, 170
– C –
concept of abstraction 45
— algorithm abstraction 50, 80
— analysis abstraction 145
— object attribute abstraction 156, 179
… …
– S –
symbol block 110
— linear 112
— rectangular 121, 167
— lowercase 130

… …
– Б –
бинарный файл 67
бит 7
блок символов 110
блок-схема 20, 170
… …
– П –
понятие абстракции 45
понятие абстракции алгоритма 50, 80
понятие абстракции анализа 145
понятие абстракции атрибута объекта 156
прямоугольный блок символов 121, 167
… …
– С –
блок символов строчный 130

Fig. 1. Fragments of subject indexes

For educational texts written in scientific and technical domains (textbooks, manuals, tutorials, etc.) subject indexes are necessary devices. Indeed, scientific texts contain many specific terms with their definitions, and as a rule, students need to study the definitions and important contexts of term usage (explaining the corresponding scientific concepts), more than once, but without reading the full text. Regretfully, subject indexes are absent in many modern textbooks and manuals for students, especially in texts of rapidly developing scientific and technical domains. To now, automated back-of-the-book indexing is an under-investigated problem, and the high-laborious text indexing work remains mainly manual since modern word processing tools provide only technical assistance.

Among recent NLP works, relatively few papers are devoted to automating back-of-the-book indexing [5–7, 13, 14], and few subject indexing systems are known: InDoc [15] and commercial TExtract[1].The central problem of index construction is extracting single-word and multi-word terms by applying linguistics and statistic criteria and filtering the more appropriate ones among extracted terms.

The works [5, 6] address the problem of extracting and filtering terms from a given text document; the proposed methods use some linguistic features of terms, statistical measures based on word occurrences, as well as machine learning, gaining precision and recall about 27–28%. The main difficulty of term detection for subject indexing relates with the fact that term extraction is performed from a single text, so various statistical measures developed and applied for corpus-based terminology extraction [9, 12] perform poorly or even are not applicable. The papers [13, 15] describes methods that are mainly based on linguistic rules for term extraction, but they are poorly described and do not provide enough information about their evaluation.

In our work, we consider the term extraction problem for subject index construction in relation to educational scientific documents, which makes it possible to use linguistic information about terminological features of such texts and thereby to achieve sufficient efficiency of the developed method. Our method of term extraction and filtering relies on rules and lexico-syntactic patterns accounting for grammatical structure of multiword terms, as well as typical contexts of their usage in the texts to be processed. Besides the linguistic rules, only simple term frequency statistics is used, without involving external text resources (so measure *tf.idf* widely used in information retrieval tasks [10] is not used).

Unlike above-mention works dealing with back-of-the-book indexing of English or French text documents [5–7, 13, 14], we consider Russian texts and exploit corresponding NLP tools. Our rule-based and corpus-free approach continues the work [4], it is close to those in [2, 15], but differs in the collection of extraction rules and in the strategy of filtering terms.

To evaluate our method, we use Russian educational scientific texts of medium size, mainly on computer science. The experiments have shown rather good performance, in average up to 70–79% of precision, recall and F-measure (the combined measure of precision and recall) for term extraction and filtering, which is considerably exceeds the results of statistics-based and machine learning methods [5, 6],

[1] http://www.texyz.com/textract/

as well as scores presented in [1] for several term extraction tools (approximately 20–47% of F-measure).

Since the results of term extraction based on modern NLP techniques are not strongly precise, the resulted list of terms needs to be validated and edited by a human expert in a problem domain, so any computer subject indexing system will inevitably be semi-automatic. Another reason for editing results by the expert is related with absence of standards on structure and content of indexes, and the work of human editors may be subjective and constructed indexes may vary in content and size. Our term extraction method is built into a research prototype of computer system[2] supporting back-of-the-book indexing and providing a user with graphical interface for setting parameters of the method and for editing results.

To clarify specialty of our approach, we begin with a brief explanation of term detection methods and their application for tasks close to subject index construction.

2. Related studies

Automated extraction of terms from texts is well investigated over last three decades. Shallow syntactic analysis along with statistical and linguistics criteria are used, based on assumption that terms are frequently encountered within texts in specific grammatical forms [2, 9]. The elaborated extraction techniques do not guarantee extracted units to be true terms (in particular, a phrase of general lexicon like *main question* may be extracted), so resulted units are considered as *term candidates* and need to be filtered. The filtering task is usually performed by evaluating and ranking the extracted term candidates with certain statistical measures and machine learning (see [12, 16]).

It should be noted that developed methods and techniques are mainly intended for extracting terms from specialized text corpora, aiming to compile terminology dictionaries or to construct thesauri and ontologies in particular domains. For these tasks, the methods have acceptable quality, but for processing single texts their effectiveness is not sufficient. Term recognition in single texts is often needed for keyword extraction [11], glossary construction [1], as well as for back-of-the-book indexing. For these tasks term extraction methods need to be modified and evaluated.

The methods developed in [5, 6] for back-of-the-book indexing rely on some grammar patterns of terms and various statistical term features based on word occurrences, but even applying machine learning they achieve about 27–28% of precision and recall. So whether the machine learning and pure statistical approach is a good choice for subject indexing seems questionable.

The works [2, 15] exploit linguistic rules for term extraction, which specify various grammatical structures of multi-word terms and their text variants encountered in the text. These rules were elaborated for corpus-based terminology extraction (from texts in French and English), and their performance for the back-of-the-book indexing task is not indicated.

[2] https://github.com/ivanov-kir-m/SISTool

The recent paper [1] describes term extraction method developed specifically for glossary construction for software requirements documents. The method uses grammatical patterns of terms along with clustering extracted terms based on certain syntactic and semantic similarity measures. In experiments with three particular software requirements documents, the method gives 35–67% of F-measure (with precision 21–51%, and recall about 90%) and slightly exceeds the best results of five term extraction tools taken for comparison. We should note that high recall and low precision is the common situation for most term extraction methods.

One can notice that the task of term extracting is quite similar to keywords recognition, but there is some difference, since terms denote concepts of a problem domain, while keywords represent main topics of the document (and may be non-terms, such as *economic trends*). However, the widely-used extraction techniques are applied for keyword extraction, and the best scores achieved on known datasets and reported in [8] are 35% of precision, 66% of recall, 45.7% of F-measure.

In contrast to the considered works, for reliable term extraction, we use a representative set of linguistic rules with lexico-syntactic patterns accounting for term features in Russian scientific texts. The formal rules are written in LSPL language [3], and the developed method has been implemented with the aid of LSPL programming tools[3].

3. Term Candidate Extraction

For extracting terms from a given text, the set of LSPL rules[4] were elaborated, based on lexico-syntactic patterns from [4]. The set encompasses three groups.

The first group of 12 rules specifies extraction of one- and multi-word terms by their typical grammatical structure (it is commonly-used by most term extraction methods [9]). The rules fix a part of speech of words (POS) and their grammatical characteristics (case, gender, etc.), for example, the pattern *N1 A N2<c=gen>* (*элементы двоичной арифметики*—elements of binary arithmetic), where *N1* is a noun, *A* is an adjective, *N2* is a noun in genitive case.

The second and the third groups of rules specify term extraction from typical contexts of term occurrences, primarily, contexts of term definitions. Such contexts are often encountered in educational scientific text, for example: *"An integrated lights-out we call remote management feature"*. Evidently, defined terms belong to significant terms to be included in the subject index.

The second group contains 53 rules for extracting terms from their definitions, covering most of the typical Russian-language phrases-definitions of terms. The rules include both particular lexical units (verbs *называть*, *определять*—*call*, *define*, and so on) and a special auxiliary pattern *Term* denoting phrase with grammatical pattern specified in the first group of rules. For example, the definition phrase *"Интегрированной средой будем называть..."* (*We call the integrated environment…*) is described by the rule:

Term <c=ins> «будем» «называть» => # Term

[3] http://lspl.ru/

[4] https://github.com/ivanov-kir-m/SISTool/tree/master/Patterns

where *Term* should be in instrumental case (*c=ins*) and is extracted (=>) in normal form (*#Term*).

The third group consists of 25 rules specifying typical contexts for introducing terminological synonyms and abbreviations in Russian scientific texts, for example: *"… информационная система, или просто ИС"* (*... information system, or simply IS ...*). The rules recognize and extract pairs of term synonyms (they should have valid grammatical patterns), relying on commas and lexical markers (e.g., words *или просто*), in the following rule the word *просто* is optional:

Term1 "," "или" ["просто"] Term2 =text> # Term1 "-" #Term2

As a result of all the extraction stage, three sets of term candidates are formed: M_{gram}, M_{def}, M_{syn}, respectively.

We have estimated the precision of term extraction for each group of rules. For this purpose two educational textbooks of medium-size on programming languages Lisp and Refal (112 and 95 pages respectively), together with their human-made subject indexes were used. For the first group of rules, experiments have shown high recall of term extraction but low precision (about 8–10%), which was expected. On the contrary, rules of the second group demonstrate high precision (90–95%) overall, due to lexical markers used in them. For similar reasons, the third group of rules shows a rather good precision: 63–67%.

Since rules and patterns of term definitions (from the second group) vary in precision, we have selected a subset of very-high precision rules, their extracted terms are labeled as *Trusted*. This label is used in our filtering procedure aiming at selection of the most important terms with the high degree of reliability.

4. Term Filtering

Based on the results of several experiments with the output sets of extracted terms M_{gram}, M_{def}, M_{syn}, we elaborated a heuristics filtering procedure that encompasses three stages.

At the first filtering stage, pre-compiled lists of stop words[5] are used. The first stop list contains words that cannot be terms (e.g., *метод, начало, отмена*—Eng.: *method*, *start*, *cancel*), while the second list contains words that cannot be part of terms, they are mainly adjectives (e.g., *данный, известный*—Eng.: *given*, *known*). From all the sets M_{gram}, M_{def}, M_{syn}, their elements are excluded that a) are encountered in the first stop list; b) contain words from the second list; c) consist of words from the first stop list. Thereby many collocations of the common scientific lexicon with the similar grammatical structure (e.g., *simple method*, *given scheme*) are discarded.

At the next filtering stage, the frequency of occurrences for all term candidates is calculated, and for frequencies of elements from M_{def}, the percentiles are calculated with the levels p_1=0.4 (rounding down) and p_2=0.95 (rounding up), respectively.

The third stage intended to account for several factors of term candidate importance: frequency of term occurrences, usage in headings/subheadings of document

[5] https://github.com/ivanov-kir-m/SISTool/tree/master/Dictionaries

sections, as well as lexical similarity of terms (that is, they have common words, e.g., *tail recursion* and *high order recursion*). According to Zipf's law, the most significant terms are units with an average frequency, and the usage of percentiles makes it possible to eliminate unlikely term candidates (both rare and frequent).

The resulting set R of subject index terms is incrementally formed according to following steps (initially R is empty).

Term candidates from the set M_{def} labeled as *Trusted*, whose frequency is in the range $[p_1,p_2]$, are added to the set R.

Term candidates from the set M_{gram}, whose frequency is in the range $[p_1,p_2]$ are added to R, provided they are encountered in some heading or subheading of the processed document (if any).

Term candidates from the set M_{gram}, whose frequency is in the range $[p_1,p_2]$, are added to R, provided they have common words (at least one) with any *Trusted* term, whose frequency is out of the range $[p_1,p_2]$.

Remaining term candidates from the set M_{def} (unconsidered in step 1) having common words (at least one) with any element from current R are added to R.

Term candidates from the set M_{def} or M_{syn}, which are synonymous to a term from R, are added to R.

All pairs of synonyms from the set M_{syn}, whose overall frequency is in the range $[p_3,p_4]$ for percentiles with levels $p_3=0.35$ and $p_4=0.95$, calculated for overall frequencies of synonymous pairs, are added to R.

Term candidates from the set M_{gram} with frequency in the range $[p_1,p_2]$ are added to R, provided they have common words (at least one) with an element from current R.

The order of the steps was determined experimentally, as well as the levels of percentiles (p_1,p_2,p_3,p_4), but the levels may be regarded as parameters be changed by a user of the subject indexing system.

5. Experiments and discussion

The encountered problem for performing experiments is the lack of human-built indexes in many Russian educational texts (textbooks, tutorials, etc.) available in electronic form (whereas many printed books have them). So we have performed experiments with 5 medium-sized (about 20 thous. words) tutorials taken from the educational resourse[6]: they are devoted to programming languages (PL), programming systems (PS), heuristic search methods (HS) in artificial intelligence. All these textbooks contain back-of-the-end indexes constructed by their authors, we regarded them as etalon sets of terms and evaluated the quality of our term filtering procedure by recall, precision, and F-measure. For comparison, we also have processed and evaluated the manual devoted to academic writing (AW), since it can hardly be attributed to scientific or technical text. The results of the evaluation are shown in Table 1.

While measuring precision and recall we had to account cases when formally different term candidates denote the same concept, for example: *условная конструкция — условие*; *conditional construction—condition*)—we considered them as term variants.

[6] http://al.cmc.msu.ru/node/4

Our filtering procedure significantly reduces the set of extracted term candidates, leaving in average about 8% of the terms. For 5 scientific texts, recall proved to be from 0.72 to 0.84, while precision varies from 0.56 to 0.77. The recall is sufficient for constructing subject indexes, and precision is acceptable, as well as F-measure. The low recall obtained for the manual on academic writing (the last row of the Table 1) is partially explained by lack of explicit definitions of certain important but relatively rare used terms (e.g., *аннотация—abstract*).

Table 1. Recall and Precision of term extraction and filtering

Text	Size (in words)	Number of terms		Precision (P)	Recall (R)	F-measure
		Extracted	Selected			
PL1	21,060	1,591	140 (8.80%)	0.74	**0.84**	**0.79**
PL2	14,322	1,012	169 (16.70%)	0.56	0.82	0.67
PL3	21,376	1,612	77 (4.78%)	**0.77**	0.72	0.75
HS	19,471	1,806	98 (5.43%)	0.71	0.74	0.73
PS	25,526	3,372	208 (6.17%)	0.70	0.81	0.75
Mean	**20,351**	**1,879**	**138 (7.34%)**	**0.70**	**0.79**	**0.74**
AW	11,699	1,884	67 (3.56%)	0.72	0.55	0.62

Our analysis of detected cases of incompleteness and inaccuracy of term extraction shows that the main reason relates to restrictions of the applied linguistic rules and lexico-syntactic patterns. In particular, certain terms are not extracted because of their complex or unusual grammatical structure (e.g., term *поиск вглубь* with pattern *N +Adverb*), which is not represented in the current collection of patterns. We also found in the texts complex phrases (with ellipsis) that define at the same time several terms, and corresponding phrase patterns are absent now. Another reason for low recall is incorrect tokenization of terms with hyphens and non-letter symbols (such as *И/ИЛИ-граф—AND/OR graph*), which leads to loss of the terms.

The analysis also shows that some extracted terms absent in the etalon subject indexes (such as term *logic programming* from the manual on Prolog) are terms relevant for indexing, and they may be omitted by human indexer because of his/her subjectivity or intent to get a more short index. So, in subject indexing task recall is more crucial than precision (provided that the number of extracted terms is not too large), since for human editor it is easier to discard some terms than to add new ones to subject index being constructed.

Overall, our method of term extraction and filtering considerably increases precision and recall in comparison with the known statistics-based methods [5, 6] and it also outperforms F-measure of the method [1]. At the same time, there are perspectives to improve the quality of term extraction, in particular, by accounting for more complex patterns and refinement of text tokenization.

Taking in mind that precision may depend on the size of processed text (the larger is text, the more terms are extracted), we have performed another experiment. Two texts (PL1 and HS) were divided into their section (chapters), which were processed and evaluated separately—the results are given in Table 2. The rows *Total*

index contain scores for total indexes obtained after merging term sets extracted separately. One can notice that for the first text (PL1) the separate processing and merging give worse F-measure (0.64 instead of 0.79), but for the second one (HS), F-measure is slightly better (0.75 instead of 0.73), and in both cases recall increases. Therefore, the strategy of separate indexing of text sections and merging of resulted indexes seems reasonable (when sections are conceptually relatively independent) and may be chosen by a user of the indexing system.

Table 2. Evaluation of merging terms extracted from text sections

Text	Section	Size (words)	Number of terms		Precision (P)	Recall (R)	F-measure
			Extracted	Selected			
PL1	1	9,886	803	50 (6.23%)	0.83	0.71	0.76
	2	5,573	329	14 (4.25%)	0.58	0.40	0.47
	3	4,880	593	314 (52.95%)	0.42	0.87	0.56
	4	6,907	426	75 (17.60%)	0.67	0.83	0.74
	Total index				**0.50**	**0.89**	**0.64**
HS	1	4,150	523	37 (7.10%)	0.77	0.69	0.73
	2	10,853	1,062	100 (9.41%)	0.58	0.70	0.63
	3	4,468	536	23 (4.30%)	0.75	0.80	0.77
	Total index				**0.72**	**0.79**	**0.75**

6. Conclusions and Future Work

We have proposed and described the term extraction method for constructing back-of-the-book index of a given educational scientific document in Russian. The method was experimentally evaluated, it demonstrates quite good performance (in average, 70–79% of F-measure) exceeding the widely-used standard methods, mainly due to the rules and lexico-syntactic patterns representing specific term usage in educational scientific texts. Thus, perspectives of rule-based methods for subject index construction of single documents seem encouraging.

The described method is implemented (with the aid of C# programming language) in a research prototype system supporting index construction. The user of the system can set parameters of the method, as well as indicate text fragment to be processed, and then verify and edit the results.

In order to accomplish more accurate and complete term extraction for subject indexing task, we evidently need to perform more experiments with texts. Future research directions are following:

- To elaborate additional lexico-syntactic patterns, in particular, patterns of non-standard phrases of term definitions;
- To improve the filtering procedure by experimenting with its parameters and the order of its steps;
- To develop methods for detecting and clustering synonymous variants of terms.

References

1. *Arora, C., Sabetzadeh, M., Briand, L., Zimmer, F.* (2016), Automated Extraction and Clustering of Requirements Glossary Terms, IEEE Transactions on Software Engineering, Vol.43, Issue 10, pp. 918–945.
2. *Aubin, S., Hamon, T.* (2006), Improving Term Extraction with Terminological Resources, Advances in Natural Language Processing, Springer Berlin Heidelberg, pp. 380–387.
3. *Bolshakova, E., Efremova, N., Noskov, A.* (2010), LSPL-Patterns as a Tool for Information Extraction from Natural Language Texts, New Trends in Classification and Data Mining, ITHEA, Sofia, pp. 110–118.
4. *Bolshakova E. I., Efremova N. E.* (2015), A Heuristics Strategy for Extracting Terms from Scientific Texts, Analysis of Images, Social Networks and Texts. Fourth Int. Conference AIST, CCIS, Vol. 542. Springer Berlin Heidelberg, pp. 285–295.
5. *Csomai, A., Mihalcea, R.* (2007), Investigations in Unsupervised Back-of-the-Book Indexing, Proc. of the Florida Artificial Intelligence Research Society Conference, pp. 211–216.
6. *Csomai, A., Mihalcea, R.* (2008), Linguistically Motivated Features for Enhanced Back-of-the Book Indexing, Proceedings Annual Conf. of the ACL, ACL/HLT, Vol. 8, pp. 932–940.
7. *Da Sylva, L.* (2013), Integrating Knowledge from Different Sources for Automatic Back-of-the-Book Indexing, Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI.
8. *Hasan, K. S., Ng, V.* (2014), Automatic keyphrase extraction: a survey of the state of the art, Proceedings of the 52th Annual Meeting of the ACL, pp. 1262–1273.
9. *Korkontzelos, I., Ananiadou, S.* (2014), Term Extraction. In: Oxford Handbook of Computational Linguistics (2nd Ed.), Oxford University Press, Oxford.
10. *Manning, C. D., Raghavan P., Schütze H.* (2008), Introduction to Information Retrieval, Cambridge University Press, New York, pp. 405–416.
11. *Matsuo, Y., Ishizuka, M.* (2004), Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information, International Journal on Artificial Intelligence Tools, 13 (1), pp. 157–169.
12. *Pecina, P., Schlesinger, P.* (2006), Combining Association Measures for Collocation Extraction, Proc. of the COLING/ACL 2006 Main Conference Poster Sessions, pp. 651–658.
13. *Reinholt, K., Lukon, S., Juola, P.* (2010), A Machine-Aided Back-of-the-Book Indexer, Proceedings of DHCS 2010, Chicago, Illinois.
14. *Wu, Z. et al.* (2013), Can Back-of-the-Book Indexes be Automatically Created? Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. ACM, pp. 1745–1750.
15. *Zargayouna, H., El Mekki, T., Audibert, L., Nazarenko, A.* (2006), IndDoc: an Aid for the Back-of-the-Book Indexer, The Indexer, 25(2), pp. 122–125.
16. *Zhang, Z., Iria, J., Brewster, C., Ciravegna, F.* (2008), A Comparative Evaluation of Term Recognition Algorithms, Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), pp.2108–2111.

USING MACHINE TRANSLATION FOR AUTOMATIC GENRE CLASSIFICATION IN ARABIC

Bulygin M. V. (bulyginmv1996@gmail.com)[1],
Sharoff S. A. (s.sharoff@leeds.ac.uk)[1,2]

[1]Russian State University of Humanities, Moscow, Russia
[2]Leeds University, Leeds, UK

This paper addresses the task of automatic genre classification for Arabic within the Functional Text Dimensions framework, which allows texts to get a reliable genre description, while maintaining an adequate amount of genre labels. Our aim in this study is to build an automatic classification model that can annotate any Web text in Standard Arabic in terms of genres. To build the training corpus we translated English and Russian annotated texts into Arabic using Google MT. For building the model experimented with various machine learning approaches, such as Logistic Regression, SVM, LSTM, and different features, such as words, character n-grams and embedding vectors. For testing the classification models, we collected and annotated in terms of FTDs our own corpus of Arabic Web texts. The best performing model offers reasonable classification accuracy in spite of being based on a training corpus produced by MT.

Key words: Functional Text Dimensions, genre classification, machine translation, Web corpora annotation

ИСПОЛЬЗОВАНИЕ МАШИННОГО ПЕРЕВОДА В ЗАДАЧЕ АВТОМАТИЧЕСКОЙ ЖАНРОВОЙ КЛАССИФИКАЦИИ ДЛЯ АРАБСКОГО ЯЗЫКА

Булыгин М. В. (bulyginmv1996@gmail.com)[1],
Шаров С. А. (s.sharoff@leeds.ac.uk)[1,2]

[1]Российский государственный гуманитарный университет, Москва, Россия
[2]Университет Лидса, Лидс, Великобритания

1. Introduction

Understanding genre is essential for processing and comprehending different kinds of texts. One way to define a genre is as "a set of conventions that transcend individual texts, and create frames of recognition governing document production, recognition and use." [Santini et al., 2010] In our lives we encounter all sorts of texts, and our ability to classify them as different genres makes working with them easier. By grouping texts together, we can identify some regularities; this, in turn, helps us understand their communicative purpose and context. From this we can build our expectations of this text, our reaction to it and our response. Automatic genre classification using methods of computational linguistics is especially useful when we are talking about the overwhelming amount of texts from the web.

The Arabic language presents some issues for the researchers in the field of automatic genre classification. First, due to Islam playing an integral part in the Arabic community, we see many more texts about religion in Arabic data as opposed to data from other languages. In our research we are avoiding this disbalance when collecting texts for test annotation. We do this in order to get a well-rounded testing of our model. Another issue that a researcher working with Arabic faces is the dominance of Arabic dialects in informal communication. According to Ethnologue, there are 36 variations of the Arabic language [Ethnologue, 2015]. They differ from each other substantially, and thus cannot be treated as the same language. Standard Arabic is the language that unites all regions of the Arab world; however, no one acquires it as his or her mother language. Because Arab children typically learn Standard Arabic in schools (which not every child has the opportunity to attend [Vasil'ev, p.c.]), it is not spoken by the entire population. On social media Arabs tend to write in their regional dialects, because these constitute the language of communication with friends and family in daily life. For our model we used texts only in Standard Arabic.

There are many different systems of distinguishing between genres. Some of them tend to present long lists of genre labels that cover all possible varieties of texts. For example, [Görlach, 2004] classifies texts into 2100 different genres and [Adamzik, 1995] presents a list of over 4000 genres. These classifications are aimed at covering all of the different possibilities that exist in the world, but they are impractical for corpora purposes. Corpora require a smaller number of genre labels to collect reasonably sized subcorpora and to make it possible to compare language use between them [Sharoff, 2018]. Long lists are also not sensitive to genre hybridism. This is even more significant for Web texts, where boundaries between genres are not strict and people can blend them or violate certain conventions.

In this work we are using the FTD (Functional Text Dimension) approach for genre classification. It provides coverage power that is similar to that of long list genre systems, while maintaining an adequate amount of genre labels. It is also much more flexible and sensitive to genre hybridism. This is possible due to the description of a text's genre through a combination of several parameters at the same time. Thus, a text may receive not only typological, but also topological analysis [Sharoff, 2018]. The texts are distinguished among 18 Functional Text Dimensions, which represent a functional category instead of an atomic label, as in other genre classification

systems. Whether a text belongs a given functional dimension is decided by a system of key questions. The texts can be scored on the following scale:

FTD value	The level of pertaining to the FTD
2.0	Strongly
1.0	Somewhat or partly
0.5	Slightly
0.0	None or hardly at all

With "0" or "None" being the default value for a FTD. The non-zero FTD values are assigned judging on how close the text is to a prototypical representative for this FTD. In our research we treated the "0.5" FTD value like the "0" score. That is why, the "0.5" score is not represented in evaluation.

2. Data

In this study we used a collection of annotated texts prepared by [Sharoff, 2018]. This collection consists of texts from 3 corpora: 5g—the Pentaglossal corpus [Forsyth and Sharoff, 2014], ukWac [Baroni et al., 2009] and GICR [Piperski et al., 2013]. 5g presents a set of texts coming from fiction, political debates, TED talks, etc. For our experiment we collected texts in English from the 5g corpus. ukWac consists of Web texts and news in English, which were collected by crawling the .uk domain. The GICR corpus represents a variety of genres such as news, articles from Wikipedia and several blogging platforms, collected from the Russian Web.

We used the method of Machine Translation to translate these annotated corpora to Standard Arabic. The resulting corpus was used as training data for our model (see Table 1)

Table 1: Size and composition of the training corpus

Corpus	Documents	Words
5g	247	686,568
ukWac	257	179,235
GICR	806	837,868
Total	**1,310**	**1,703,671**

We used Google Translate for translation, because it shows good performance and is accessible to any user. The quality of translation was quite high, because Google Translate managed to preserve most grammatical relations and syntactic boundaries. The most common mistake was inadequate word choice. For example, the Russian sentence "Спасибо, только день рождения был более полугода назад" ("Thank you, except the birthday was more than half a year ago") was translated to the following:

(1) shukran eid milad faqat kan ‘akthar min nisf eam
THANK YOU HOLIDAY OF BIRTH ONLY WAS MORE THAN HALF YEAR

However, we would expect to see this:

(2) shukran lakin eid milad kan ‘akthar min nisf eam
THANK YOU BUT/ONLY HOLIDAY OF BIRTH WAS MORE THAN HALF YEAR

In our research we decided to conduct an experiment to find out whether a model can be trained on a corpus translated with Machine translation. The resulting program would be judged by how well it can classify natural Arabic texts.

For the testing set of data we collected and annotated 100 Arabic texts from the Web. There were 24 different sources: news, fiction, Wikipedia, scientific texts, law texts, etc. We did not use texts from Facebook or Twitter, because people do not use Standard Arabic there. The same issue takes place with forum dialogs and simple discussions. Text length varies between 300 and 1500 words. Each text was annotated in terms of Functional Text Dimensions, with all principal dimensions being scored. Statistics for the testing data set are presented in Table 2 with overall sum of annotations and mean value for each dimension.

Table 2: Distributions of genres in testing data set in terms of FTDs

FTDs	A1	A3	A4	A5	A6	A7	A8	A9	A11
Total	42	7	16	2	0	35	39	10	10
Mean	**0.42**	0.07	0.16	0,02	0	0.35	0.39	0.1	0.1
FTDs	**A12**	**A13**	**A14**	**A15**	**A16**	**A17**	**A18**	**A19**	**A20**
Total	5	0	10	1	45	15	0	0	0
Mean	0.05	0	0.1	0.01	**0.45**	0.15	0	0	0

3. Experiments

3.1. Features

In this study we conducted several experiments with different features and methods for classification. For the feature selection testing we chose words and character trigrams. Genre information is strongly concentrated around word choice. Stylistic differences are often key to genre identification. To vectorize word features we used tf-idf technic [Pedregosa et al. 2011], which is one of the most common methods for text vectorization. To vectorize texts for the LSTM neural network we used pre-trained word vectors from the fastText database [Bojanowski et al. 2016]. These vectors with a dimensionality of 300 were trained on the Arabic sector of Wikipedia, using the skip-gram model.

As mentioned in [Zhang et al. 2015], the character level of a text can be very useful for text classification. For our research we used character trigrams. Text tokenization was done using scikit-learn preprocessing tools. Trigrams were also vectorized using tf-idf technic.

3.2. Methods

For each of the features we used the following methods:

1. SVM
2. Logistic regression
3. XGboost
4. LSTM

The Support Vector Machine is one of the most popular methods for learning algorithms. It is used for various tasks of machine learning, such as sentiment analysis, language modeling and text classification. It is also useful for multiclass classification as shown in [Hou et al. 2015]. In our research we set the C value of the SVM model to 1, and we used the one-versus-rest scheme for the training of the classifier. The Support Vector Machine can also learn using different kernel types: linear kernel, RBF kernel and polynomial kernel. For our model we used linear kernel.

Logistic regression is a basic method of machine learning borrowed from field statistics. It utilizes logistic function to predict the probability of an answer. It is one of the most popular methods for binary classification, but one can also use it for multiclass classification, for example, in tasks of image classification or spam filtering. The C parameter of our Logistic regression model was set to 1. We also used the one-versus-rest scheme to train the classifier.

XGBoost (Extreme Gradient Boosting) is an implementation of gradient boosted decision trees designed to increase productivity and performance. It is an open-source software library [Chen, Guestrin, 2016]. This algorithm became widely used recently because it has been dominating applied machine learning tournaments and Kaggle competitions. The key advantages of this method are its speed and accuracy compared to similar methods. It is highly flexible and versatile for most machine learning tasks, such as classification, regression or ranking problems. For our model we used 300 estimators and we set the learning rate to 0.05.

LSTM (Long short-term memory) is a deep-learning technique. This model is a Recurrent Neural Network with a special architecture that allows it to avoid the problem of vanishing gradient and to learn long-term dependencies. The latter is especially useful for the task of text learning. We chose LSTM for our model because it is less dependent on a big training corpus compared to other popular neural network architectures. However, in order for LSTM to achieve good results a big training corpus is still needed. In our work we encountered this problem by trying to train the LSTM model on texts simply vectorized through the tf-idf technique. Because our corpus is relatively small, the training was not successful. In order to compensate for this, we implemented semantic vectors from the fastText database [Bojanowski et al. 2016]. These vectors also include information about the inner structure of the word. This was done by training the model using character n-gram features. For our LSTM network we used the Adam optimizer. Our model trained for 500 epochs with batch size 14.

4. Evaluation

Although overall accuracy is the easiest and most intuitive metric for model evaluation, it is quite useless for the purposes of evaluating multiclass classification. So instead we used precision, recall and F_1-score metrics. These metrics allowed us to understand how well our model can detect different genres and how good it is at distinguishing genres between each other. We also used the F_1-score metric that summarizes the results of precision and recall evaluation.

We tested our model with all of the methods and features described in the previous section. For overall results of the performance of all models see Table 3.

Table 3: Different classifiers' overall performance

Classifier + Feature	Precision	Recall	F_1-score
SVM + Words (tf-idf)	0.91	0.93	0.91
Logistic regression + Words (tf-idf)	0.90	0.93	0.91
XGBoost + Words (tf-idf)	0.90	0.93	0.91
LSTM + fastText vectors	0.87	0.88	0.87
SVM + Character trigram (tf-idf)	0.92	0.94	0.92
Logistic regression + Character trigram (tf-idf)	0.90	0.93	0.91
XGBoost + Character trigram (tf-idf)	0.91	0.94	0.91

As the result of the evaluation of our model, we see that the three best performing models are the SVM classifier trained on character trigrams, the XGBoost classifier also trained on character trigrams and the SVM trained on words. We also tested these three models for the performance of classification for each value that can be assigned during the classification in terms of FTD. In Tables 4–6 we compare the collective result for each FTD value across all 18 genres.

Table 4: The performance of the SVM + character trigram model for each FTD value

FTD value	Precision	Recall	F_1-score
0	0.94	0.99	0.97
1	0.22	0.07	0.11
2	0.72	0.27	0.39

Table 5: The performance of the XGBoost + character trigram model for each FTD value

FTD value	Precision	Recall	F_1-score
0	0.94	1.00	0.97
1	0.00	0.00	0.00
2	0.76	0.18	0.29

Table 6: The performance of the SVM + words model for each FTD value

FTD value	Precision	Recall	F_1-score
0	0.94	0.99	0.96
1	0.29	0.07	0.11
2	0.57	0.16	0.25

We also tested how well our best performing model can classify each genre. For each of the 18 Functional Text Dimensions we computed overall precision, recall and F_1-score. Each overall score is calculated accordingly to the share of each FTD value.

Table 7: The performance of the SVM + character trigram model for each FTD

FTDs	A1	A3	A4	A5	A6	A7	A8	A9	A11
Precision	0.79	0.96	0.96	0.98	1.00	0.84	0.84	0.96	0.87
Recall	0.79	0.95	0.96	0.97	0.98	0.85	0.87	0.96	0.93
F_1-score	0.72	0.95	0.95	0.97	0.99	0.81	0.85	0.95	0.90
FTDs	**A12**	**A13**	**A14**	**A15**	**A16**	**A17**	**A18**	**A19**	**A20**
Precision	0.90	0.99	0.88	0.98	0.72	0.85	1.00	1.00	0.99
Recall	0.95	1.00	0.94	0.97	0.77	0.92	0.98	0.98	0.98
F_1-score	0.93	0.99	0.91	0.97	0.72	0.88	0.99	0.99	0.99

Our model shows great results for some FTDs. However, this can be due to the low representability in the test corpus for these FTDs. Thus, the results for FTDs that have total value less than 10 in Table 2 can be interpreted as a majority class classification.

The best overall performing model is the SVM + character trigram. It shows the best result for the overall precision evaluation and it shares the first place for the overall recall. It can identify different FTD values and it also shows good performance for the one of the most represented A8 dimension (hardnews) (see Table 8). The zeros for the "1" FTD value are explained by the low number of "1" FTD values in our testing corpus.

Table 8: The performance of the SVM + character trigram model for the A8 dimension

FTD value	Precision	Recall	F_1-score
0	0.89	0.96	0.86
1	0.00	0.00	0.00
2	0.73	0.61	0.67

Classifiers that were tested in our research are "black-box" classifiers, which means that we do not know how the parameters for classification were set. However, the SVM classifier allows us to look at the most valuable features for each class, which can tell us a lot about the classification process. We analyzed our training corpus in terms of the most valuable features for determining each FTD. We did it using a SVM model that used character trigram as features. For the A4 (fiction) FTD the

most valuable features were “قد”, “قال”, “ها”. The first two are past tense markers and can also be interpreted as the elements of a narrative. “ها” in its primary meaning is a feminine possessive pronoun, however it is also used in some complex structures, for example, relative clauses. For A7 (instruct) FTD the most valuable features are “إذا” and “إذ”. These are conjunctions that form conditional sentence. “لي”, “ني”, “نا” and “أنا” were the features with highest weights in the A11 (personal) FTD. All of these features represent different variations of the pronoun “I”. The markers of A8 (hardnews) FTD are verb “قال” (to say) and conjunction “ان” that usually follows that verb to form indirect speech. The top features of the A9 (legal) FTD were “حاد”, “تحا” and “اتح”, which form one word “اتحاد” (union).

5. Analysis of results

5.1. Quality of detection

There are several conclusions that emerge from the results of our experiment. First, all of the models are pretty good at detecting when some text is not represented in a FTD, and thus has the value of “0”. All of them also have over 90% in precision and recall metrics for this FTD value. The key factor to the good performance of the SVM and XGBoost models was that they managed to correctly detect the FTDs with the value of “2” for the majority of texts. These models achieved high results for the precision metric, but they did not perform as well for the recall metric. Thus, our models perform well for the task of distinguishing texts that belong to different FTDs, but they are not as good at detecting all texts that belong to one FTD.

The sad conclusion that comes from the results of our experiment is that no models were able to correctly identify the “1” FTD values. The best algorithm for sensing the middle values of a genre was the algorithm that used LSTM with fastText word vectors. This model also showed good results in the recall metric for the FTD value of “2”, so it could adequately identify when a text pertained strongly to a FTD. However, the overall performance of the LSTM model was rather disappointing. It did not show good results for distinguishing texts between genres, which was the reason that this model got low scores.

The best performing feature was the character trigram; this corresponds to the results in [Sharoff et al, 2010]. However, it is hard to compare the overall results of our classification with other works, because to our knowledge no such research was conducted on Standard Arabic material. The results of the classification also depend greatly on the collection of texts used in the experiment.

5.2. Common problems

The main problem of our experiment is that our models did not achieve high scores for the recall evaluation for the non-zero values. A possible reason for this may be that for some texts the best performing classifiers did not assign a “2” value for the

FTD at all, which resulted in some texts being poorly scored. It is possible that the cause of this problem is that we used translated texts for the training and then tested our models on real texts from the Web.

Another problem that we encountered was the fact that we did not have a big training corpus. A bigger training corpus would likely have increased the results of most of the classifiers that we tested. A bigger corpus is definitely required for deep-learning techniques, such as LSTM. During our research we tried to train our LSTM model by simply vectorizing words in our corpus with the tf-idf technique, but because our corpus was relatively small, we were not able to do it.

6. Conclusions

In this paper we presented an experiment in which we built a model that classifies Arabic Web texts in terms of Functional Text Dimensions. For this experiment we produced a training corpus using machine translation tools. The original corpus and its annotations were taken from [Sharoff, 2018]. We also conducted an experiment to find out what features and what classifier has the best performance in this environment. The SVM classifier with character trigrams showed the best results. For testing the models, we collected a small representative corpus from Arabic websites with texts in Standard Arabic. They were annotated in terms of Functional Text Dimensions. The best performing model achieved precision of 93% of correctly identified FTD values. This indicates that the genre features are well preserved through Machine Translation.

Further work is still necessary, as we encountered several problems throughout our research. One of the possible directions of further work concerns building a bigger training corpus with more diverse texts in terms of their functional categories. A larger corpus would allow us to experiment with deep-learning techniques, such as the different architectures of the Recurrent Neural Network and the Convolutional Neural Networks. Another interesting development for our research could be the comparison of our classification to the classification of other corpora of Standard Arabic.

Experiments with semi-supervised settings for LSTM show promising results for the task of text classification [Johnson and Zhang, 2016]. We need to implement this technic in our future work.

We also need to increase the representativeness of our testing corpus. So far, a substantial part of it can be classified using only 3–4 dimensions. We need to collect and annotate more texts from personal blog-entries, opinion columns, product reviews, etc.

In our research we have seen that our model is capable of achieving reasonably good results. However, it still is not reliable with respect to recall by failing to detect many texts that belong to the same Functional dimension. This means that more testing and better machine learning techniques are needed.

References

1. *Adamzik K.* (1995), Textsorten—Texttypologie, Eine kommentierte Bibliographie, Nodus, Münster.
2. *Baroni M., Bernardini S., Ferraresi A., and Zanchetta E.* (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora, Language Resources and Evaluation, 43(3), pp. 209–226.
3. *Bojanowski P., Grave E., Joulin A., Mikolov T.* (2016), Enriching Word Vectors with Subword Information.
4. *Chen T., Guestrin C.* (2016), XGBoost: A Scalable Tree Boosting System, In 22nd SIGKDD Conference on Knowledge Discovery and Data Minin.
5. *Forsyth R., Sharoff S.* (2014), Document dissimilarity within and across languages: a benchmarking study, Literary and Linguistic Computing, 29, pp. 6–22.
6. *Görlach M.* (2004), Text types and the history of English, Walter de Gruyter.
7. *Hou H., Han P., Cao D.* (2015), The Application Based on Decision Tree SVM for Multi-class Classification.
8. *Johnson R., Zhang T.* (2016), Supervised and semi-supervised text categorization using LSTM for region embeddings, in ICML.
9. *Pedregosa et al.* (2011), Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825–2830.
10. *Piperski, A., Belikov, V., Kopylov, N., Selegey, V., and Sharoff, S.* (2013), Big and diverse is beautiful: A large corpus of Russian to study linguistic variation, In Proc 8th Web as Corpus Workshop (WAC-8).
11. *Santini, M., Mehler, A., and Sharoff, S.* (2010), Riding the rough waves of genre on the web, Genres on the Web: Computational Models and Empirical Studies, Springer, Berlin/New York.
12. *Sharoff S.* (2018), Functional text dimensions for annotation of web corpora, Corpora.
13. *Sharoff, S., Wu, Z., and Markert, K.* (2010). The Web library of Babel: evaluating genre collections. In Proc. of the Seventh Language Resources and Evaluation Conference.
14. *Trevilla L.* (2009), Ethnologue: Languages of the World.
15. *Zhang X., Zhao J., LeCun Y.* (2015), Character-level convolutional networks for text classification, In Advances in Neural Information Processing Systems, pp. 649–657.

BOUNDARY EXPRESSION IN VERBS AND GESTURE: DIFFERENCES BETWEEN L1 AND L2 SPEAKERS[1]

Denisova V. A. (valeriia.deni@gmail.com)[1,2],
Cienki A. (a.cienki@vu.nl)[1,2],
Iriskhanova O. K. (iriskhanova@me.com)[1]

[1]Moscow State Linguistic University, Moscow, Russia;
[2]Vrij Universiteit Amsterdam, Amsterdam, the Netherlands

The notion of event boundaries is closely connected with the category of aspect. Aspectual forms show different views of "internal temporal consistuency of a situation" (Comrie 1976:3) and, consequently, construals of events in different ways. Recently scholars have started looking into the core of the aspectual distinction through multimodality, considering hand gestures. On the basis of Russian and French oral narratives produced by native speakers, we conducted a study, testing our hypothesis about the existence of direct correlation between the expression of boundaries in verbs and in gestures. Means of boundary expression regarded for Russian on the verbal level were perfective (soveršennyj vid) and imperfective (nesoveršennyj vid) verbs, and for French—passé composé and imparfait. On the kinesthetic level we distinguished between bounded gestures (i.e., involving a pulse of movement) and unbounded gestures (i.e., smooth by nature). While for French L1 we found a direct correlation between gesture boundary schemas and aspectual forms, the results for Russian L1 did not support our hypothesis. With a view to these differences between the two languages, we studied the boundedness correlation in oral narratives produced by Russians speaking French as L2 (CEFR levels B2-C1). The comparison between L1 and L2 narratives revealed a certain change of gestural patterns: the Russian speakers of French L2 used almost the same number of unbounded and bounded gestures with the perfective verb forms and more unbounded gestures with the imperfective forms, thus moving closer towards French L1 speakers' verb-gesture patterns. The use of gestures can be accounted for by a series of noise factors related to language peculiarities, the cognitive mechanism of profiling and challenges of speaking in L2.

Key words: grammatical aspect, event construal, Russian, French L2, embodiment, gesture studies

[1] This research is carried out at the Centre for Socio-Cognitive Discourse Studies of Moscow State Linguistic University and is supported by the Russian Science Foundation (grant №14-48-00067-П).

ВЫРАЖЕНИЕ ПРЕДЕЛЬНОСТИ В ГЛАГОЛАХ И ЖЕСТАХ: РАЗЛИЧИЯ МЕЖДУ ГОВОРЯЩИМИ НА РОДНОМ И НА ИНОСТРАННОМ ЯЗЫКЕ

Денисова В. А. (valeriia.deni@gmail.com)[1,2],
Ирисханова О. К. (iriskhanova@me.com)[1],
Ченки А. (a.cienki@vu.nl)[1,2]

[1]Московский Государственный Лингвистический Университет, Москва, Россия; [2]Vrij Universiteit Amsterdam, Амстердам, Нидерланды

1. Introduction

The notion of aspect has long been controversial in linguistics. Aspectual forms are believed to express different ways of construing events [Langacker 2008], conveying a whole series of semantical components: the distribution of an event in time, the viewpoint of an event from "inside" or from "outside", integrity or processuality, as well as boundedness of events [Bondarko 1971]; [Comrie 1976]; [Maslov 2004].

In the present study we consider the way boundedness comes through in the Russian and French aspectual systems. The two systems have points of convergence as well as differences. In general, both systems are traditionally regarded as an opposition between perfective and imperfective forms. However, in Russian perfectivity and imperfectivity are expressed on the grammatical level in the infinitive of verbs and go through the whole Russian verbal paradigm, while in French the two opposed forms exist for each verb in terms of different past tenses for the verb.

Aspect in Russian is not a purely grammatical category. Aspectual differences are also expressed at the lexical level as the meaning of the verbs in context can carry certain characteristics that are traditionally related to boundedness [Padučeva 2010]. Although French aspect is considered to be a grammatical category, lexical features are sometimes also taken into account [Garey 1957].

Recently, researchers have started considering aspect in relation to co-speech gestures (e.g. [Becker et al. 2011]; [Duncan 2002]; [Parrill et al. 2013]). Building on [McNeill's 1992] idea of the growth point as a starting point for an idea that further unfolds into complexes of gestures and language forms, scholars usually analyze gestures that are synchronized in time with verbs. The most important result is the proof that aspectuality can be studied not only from the grammatical point of view, but as a broader cognitive phenomenon crucial for the construal of events.

The present study is the continuation of our L1 research based on the analysis of the narratives produced by Russian and French L1 speakers. At the L1 stage we considered the correlation between aspectual verb forms and some of the characteristics

of the co-verb gestures. On the verbal level, perfective and imperfective verbs were opposed, while on the kinesthetic level we investigated gestures that involved a clear pulse of effort (bounded gestures) and those with smooth movements, exhibiting controlled movement (unbounded gestures) [Laban, Lawrence 1974]. We hypothesized that perfective verb forms would more often co-occur (i.e. fully or partially synchronize) with bounded gestures (reflecting an event as a whole, "in one go"), while imperfective verb forms—with the unbounded ones (relating to the sustained focus on the internal constituency of the event).

The hypothesis was tested on oral narratives about personal experience produced by native speakers of Russian and French. The quantitative results of the French L1 narratives confirmed the initial hypothesis: 105 imperfective verbs out of 177 were used with unbounded gestures, 107 out of 150 perfective verbs—with bounded gestures. The difference is significant for both, imperfective ($X^2 = 17.89$, df = 1, $p < 0.01$, $p = 0.3$) and perfective verbs ($X^2 = 27.31$, df = 1, $p < 0.01$). The Russian speakers used significantly more bounded gestures with both aspectual forms. 149 out of 237 imperfective verbs ($X^2 = 15.7$, df = 1, $p < 0.01$) and 130 out of 178 perfective verbs ($X^2 = 37.78$, df = 1, $p < 0.01$) were used with bounded gestures (see Cienki, Iriskhanova (in press) for detail). Quantitative results obtained from L1 narratives were explained by the difference between the Russian and French aspectual systems as in the Russian language aspect is closely related to lexical semantic characteristics of the verbs and is often sensitive to factors other than the grammatical opposition of perfectivity/imperfectivity.

The difference between Russian L1 and French L1 suggests that advanced French L2 speakers whose native language is Russian will gesture in a distinctive way. On the one hand, they have probably adopted the image of events expressed by verbs in *passé composé* as bounded complete entities and associate *imparfait* with ongoing and incomplete events. On the other hand, they may still follow the aspectual differences in Russian in which lexical meanings of verbs play an important role and in some contexts outweigh the neat division between grammatical perfectivity and imperfectivity.

1.1. Methodology

Relying on the results of the L1 study, we conducted a similar study for L2 speakers: 22 Russians who had acquired French to an advanced level (B2-C1, CEFR) produced oral narratives of personal experiences, using the protocol of [Becker et al. 2011]. The narratives were videotaped and annotated for aspectual forms and gesture characteristics. The participants took part in the experiment in pairs. Each pair sat opposite a camera and a laptop, with the assignments for the experiment displayed on the screen. The subjects were asked to tell stories to each other on certain topics. All the topics from the list were displayed on the screen and encouraged the participants to produce a story about a real event, in which they had either participated or which they had witnessed themselves. The length of each video was approximately 12 minutes. Each video contained at least six narratives on the proposed topics.

2. Analysis of the narratives

For the quantitative analysis we considered the number of perfective and imperfective verbs that co-occurred with bounded and unbounded gestures in the L2 data. The data displayed the following results (see Fig. 1).

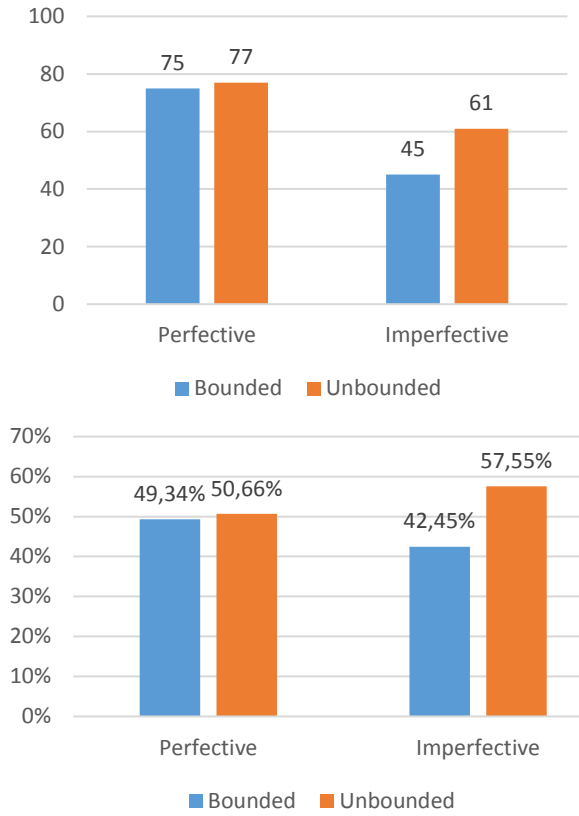


Fig. 1. Quantitative results for L2 data

This result differs from the Russian and French L1 results. The French L1 speakers tend to use more bounded gestures with perfective verbs and more unbounded with imperfective ones; the Russian L1 speakers tend to use more bounded gestures with both types of verbs. However, the Russian speakers of French used almost the same number of unbounded and bounded gestures with the perfective verbs and insignificantly ($X^2 = 2{,}415$, df = 1; $p = 0{,}12$) more unbounded gestures with the imperfective ones.

A closer look at the L2 data revealed a series of noise factors that influence the overall results for L1 and L2 speakers, which can help to explain some cases for French L1 and most cases for Russian L1 and French L2 that were inconsistent with the initial hypothesis.

2.1. Specific gesture types

Building on McNeil's growth point theory, we annotated all the gestures that co-occured with verbs, thus our datasets contained rhythmical gestures that mark the rhythm of speech, discursive gestures that mark the structure of communication, and representational gestures that illustrate the thing or the action they refer to.

The preliminary results for the Russian L1 data suggested that boundedness of events was expressed simultaneously on verbal and kinesthetic levels when gestures were representational, i.e., when they represented the semantics of verbs. It is worth mentioning that the tight link between the category of aspect and the lexical meaning of the verb in Russian has been repeatedly highlighted by numerous scholars [Bondarko 1971]; [Maslov 2004]; [Padučeva 2010].

Further analysis showed that 36 perfective verbs out of 52 co-occurred with bounded representational gestures ($X^2 = 7{,}6$, df = 1; $p = 0.008$), while 60 imperfective verbs out of 94 were used with unbounded representational gestures ($X^2 = 7{,}1$, df=1; $p = 0.007$). Thus, the results for the Russian L1 data confirmed that most representational gestures followed the initial hypothesis of the research.

The French L1 results for representational gestures follow the hypothesis with a higher statistical index than the result for all the gestures in French L1: 36 perfective verbs out of 47 were used with bounded gestures ($X^2 = 13{,}2$, df = 1; $p = 0.0003$) and 44 imperfective out of 66 were used with unbounded ($X^2 = 7{,}3$, df = 1; $p = 0.006$).

Proceeding from the results for L1 data, we assumed that representational gestures of L2 speakers will also be consistent with the initial hypothesis, and will show the direct correlation in terms of (un)boundedness in verbs and gestures. Besides, the L2 narratives were produced by Russian native speakers who might be influenced by a tighter connection between grammatical aspect and verbal semantics in their native language. In addition, most French grammar books present *imparfait* as an incomplete and ongoing action as opposed to complete and bounded *passé composé*. All these facts suggested that Russian speakers would tend to express boundedness on both verbal and kinesthetic levels when they use representational gestures to express the form of an object or embody an action.

In order to test this assumption we annotated all the representational gestures that illustrate the verbs they co-occurred with ("representational-verb" gestures). The results revealed that although there were few gestures of this type (17), practically all of them (16) confirmed our hypothesis. The small number of instances with representational gesture does not allow us to make generalized statements, nevertheless, the data indicates that in most cases the expression of boundedness in the verbs correlated with its expression in the representational-verb gestures.

2.2. Noise factors influencing the overall data

The cases when boundedness expressed in gestures did not correlate with boundedness expressed in aspectual forms present quite a controversial issue. We conducted a qualitative analysis of such cases and grouped them according to the reasons for which the correlation was absent.

The first group of factors is related to the language specific features. Inconsistencies in Russian L1 data often occurred because of the fact that some of the imperfective verbs bear the idea of boundedness within themselves, for example, *ja tuda ezdilaImperf mnogo raz (I went there many times).* Another controversial issue for Russian consists in the use of infinitive constructions, containing 2 different aspectual forms, for instance, *načaliPerf krutitsjaImperf Inf (started spinning).* In French, the perfective form (*passé composé*) is an analytical one, as it is built up out of two components: the verb *avoir* (*to have*) or *être* (*to be*) and past participle of a notional verb. Moreover, some types of adverbs are placed between the two parts of the form, e.g. *j'ai toujours pensé* (*I always thought*). Consequently, the two parts of a single form were sometimes synchronized with different gestures that might also be of different types in terms of boundedness.

The second group of factors is related to the cases in which gestures clearly represented or referred to an object, and not to the action expressed in the verb. For instance, one of the participants of the L2 study used the phrase *nous avons dessiné de grands lettres (we drew big letters)* that synchronized with a representational gesture, illustrating how big the letters were. The speaker made a broad gesture with her arms moving in opposite directions (see Fig. 2). Thus, the gesture illustrates *de grands lettres (big letters)*, even though it partially correlates with the verbal form *avons dessiné (drew).*

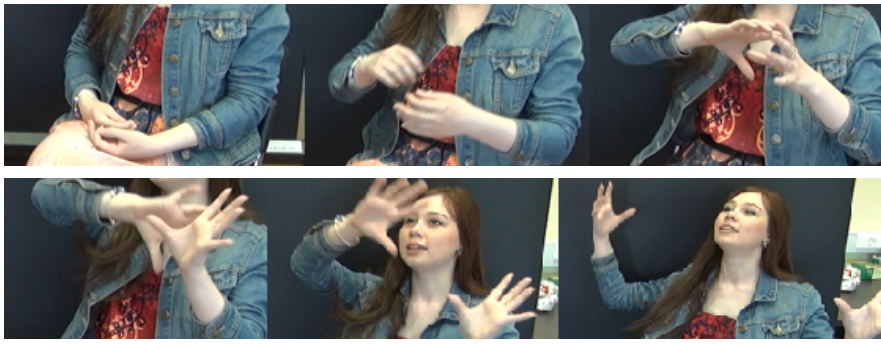


Fig. 2. Representational-other gesture

In another narrative a participant interrupted her interlocutor with the phrase *tu as choisi une table (you chose a table).* The phrase co-occurred with a pointing gesture, so the speaker highlighted the participant of the event through the gesture produced (see Fig. 3).

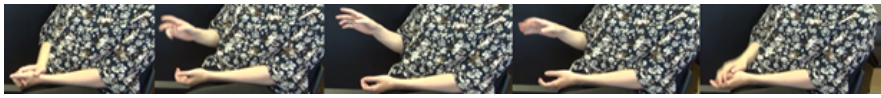


Fig. 3. Pointing gesture

Another example concerns the notion of negation. In numerous cases when a verb was used in the negative form, the negation was marked in the form of a gesture that primarily highlighted the idea of negation [Calbris 2011]. Talking about the challenges of making an order in a foreign restaurant, one of the participants said *elle ne m'a compris…du tout (she didn't understand me… at all).* The verb phrase synchronized with a hand movement that went from left to right, reminding in path of a head shaking in negation (see Fig. 4).

Fig. 4. Negation gesture

Such cases can be explained through profiling, a cognitive mechanism of construal introduced by [Langacker 2008]. In the examples mentioned above the gestures profile certain event characteristics which might not correlate with the features profiled by the verb that co-occurs with the gesture.

2.3. Challenges of L2 data

The L2 data is expected to differ from the L1 data due to the fact that L2 speakers might make mistakes, forget some collocations or grammar patterns. The participants of our study acquired French on the level which was high enough for the kind of task they were given. That is why there were few mistakes in their use of grammatical forms, however, in some cases the speakers could not recall a word for several seconds and while thinking they used cyclic movements that are traditionally called *search for word gestures* [Ladewig 2011]. Such cycling movements often co-occured with self-correction or a prolonged hesitation sound *erm (erm…on était…avait [erm …he was… he had]; j'ai commencé…erm… j'ai commencé [I started…erm…I started).*

In another example the same verb was used twice with hesitation pauses and in both cases it was accompanied by two different gestures. The participant produced the phrase *elle faisait…elle faisait le ménage (she was doing...doing the cleaning).* The verb *faire (to do)* was used two times in the form of *imparfait*, each time with a different gesture: in the first case—with an unbounded gesture, in the second case—with a bounded one. Both gestures were cyclic, with a small pause between them (see Fig. 5).

Fig. 5. Search for word gestures

The second feature that occurred in L2 data was a small number of representational gestures produced by the speakers, while in L1 data the gestures of this kind appeared to be more widespread. The fact that representational gestures were not abundant in our data can be due to the cognitive overload of L2 speakers, who tend to use non-representational gestures to facilitate lexical retrieval in L2, rather than actually illustrate the verbs with gestures to make their speech more expressive [Gullberg 1998]. Overall, representational gestures of L2 speakers reveal results similar to the ones of L1 speakers of Russian. Thus, it seems that when an event that is concisely transmitted through a verb form is illustrated by a representational gesture, its (un) boundedness is also expressed on both levels (kinesthetic and verbal).

3. Discussion

The results obtained from the L1 speakers revealed that the French L1 speakers tend to use bounded gestures with perfective verbs and unbounded with imperfective verbs, while the Russian L1 speakers tend to use more bounded gestures for both aspects. The L2 speakers demonstrated a mix of French and Russian gesture patterns. As the quantitative analysis showed, there was no direct correlation between boundary expression on the verbal and gestural levels in the Russian L1 and in the French L2 data obtained from the Russian speakers.

A possible explanation is that for the Russian L2 speakers the construal of events was not necessarily centered around the verb, but could be conceptualized holistically in the whole phrase, thus often profiling some characteristics that were expressed in some other words related to the event as a complex entity. Further investigation is needed to determine whether this feature is characteristic of other (non-Russian) L2 speakers.

Overall, the study indicates how events are construed and expressed by Russians speaking French in simulated oral narratives, so the results are applicable to a specific type of discourse and to a particular group of people. The significance of noise factors needs to be tested on a bigger corpus. However, this kind of research demonstrates that a multimodal approach to event construal in different languages can reveal the subtle nature of gesture-verb correlation in the transition from L1 to L2.

References

1. *Becker R., Cienki A., Bennett A., Cudina C., Debras C., Fleischer Z., Haaheim M., Müller T., Stec K., Zarcone A.* (2011), Aktionsarten, speech and gesture; Proceedings of GESPIN 2011: Gesture and speech in interaction, pp. 30–35.
2. *Bondarko A. V.* (1971), Tense and aspect of Russian verb [Vid i vremja russkogo glagola], Prosveščenije, Leningrad.
3. *Comrie B.* (1976), Aspect, Cambridge University Press, Cambridge.
4. *Calbris G.* (2011), Elements of meaning in gesture, John Benjamins Publishing Company, Amsterdam.
5. *Duncan S.* (2002), Gesture, verb, aspect, and the nature of iconic imagery in natural discourse, Gesture Vol. 2 (2), pp.183–206.
6. *Cienki A.,Iriskhanova O.* (Eds.) Aspectuality across languages: Event construal in speech and gesture, John Benjamins Publishing Company, Amsterdam (in press).
7. *Garey H. B.* (1957), Verbal Aspect in French Language, Language, Vol. 33 (2), pp. 91–110.
8. *Gullberg M.* (1998), Gesture as a Communication Strategy in Second Language Discourse. A Study of Learners of French and Swedish, Lund, Lund.
9. *Laban R., Lawrence F. C.* (1974), Effort: Economy of Human Movement, Macdonald and Evans, London.
10. *Langacker R. W.* (2008), Cognitive grammar: A basic introduction, Oxford University Press, Oxford.
11. *Ladewig S. H.* (2011), Putting the cyclic gesture on a cognitive basis, CogniTextes, Vol. 6, available at: http://journals.openedition.org/cognitextes/406
12. *Maslov U. S.* (2004), Selected works. Aspectology. General Linguistics [Izbrannije trudy. Aspectologija. Obščeje jazykoznanije]. Jazyki slavjanskoj kultury, Moscow.
13. *McNeill D.* (1992), Hand and mind: What gestures reveal about thought, University of Chicago Press, Chicago.
14. *Padučeva E. V.* (2010), Semantic studies. The semantics of tense and aspect in Russian. The Semantics of a narrative [Semantičeskije issledovanija. Semantika vida i vremeni v russkom jazyke. Semantika narrativa], Jazyki Slavjanskoy Kultury, Moscow.
15. *Parrill F., Bergen B. K., Lichtenstein P. V.* (2013), Grammatical aspect, gesture, and conceptualization: Using co-speech gesture to reveal event representations, Cognitive Linguistics, Vol. 24 (1), pp. 135–158.

НЕМЕЦКИЕ КОНСТРУКЦИИ С МОДАЛЬНЫМИ ГЛАГОЛАМИ И ИХ РУССКИЕ СООТВЕТСТВИЯ: ПРОЕКТ НАДКОРПУСНОЙ БАЗЫ ДАННЫХ[1]

Добровольский Д. О. (dobrovolskij@gmail.com)
Институт русского языка РАН, Москва, Россия

Зализняк Анна А. (anna.zalizniak@gmail.com)
Институт языкознания РАН; ФИЦ ИУ РАН, Москва, Россия

В статье излагаются принципы контрастивного корпусного исследования немецких и русских модальных конструкций. Ставится задача, во-первых, уточнить номенклатуру значений немецких модальных глаголов и условий их реализации, а во-вторых, выявить и описать средства выражения модальных значений в русском языке на основе анализа множества конструкций, служащих функциональными эквивалентами при переводе на русский язык конструкций с немецкими модальными глаголами. Анализ предлагается осуществлять при помощи создания на основе репрезентативного массива параллельных немецко-русских текстов Национального корпуса русского языка (НКРЯ) *надкорпусной базы данных* переводных соответствий, в которой как немецкой конструкции с модальным глаголом, так и ее русскому переводному эквиваленту приписывается аннотация в форме набора значений релевантных признаков. Такая база данных, с одной стороны, будет представлять собой ценный лингвистический ресурс, который может быть использован, в том числе, для создания нового поколения электронных интерактивных немецко-русских и русско-немецких словарей; с другой стороны, построенная на основе этой базы данных инвентаризация типов конструкций русского языка с (потенциальным) значением модальности составит важный вклад в грамматику конструкций русского языка, подтверждающий принципиальную непрерывность в отношениях между лексикой и грамматикой.

Ключевые слова: модальность, модальные глаголы, русский язык, немецкий язык, параллельный корпус, надкорпусная база данных, двуязычная лексикография

[1] Работа выполнена при поддержке РФФИ (грант 17-29-09154).

GERMAN CONSTRUCTIONS WITH MODAL VERBS AND THEIR RUSSIAN CORRELATES: A SUPRACORPORA DATABASE PROJECT

Dobrovol'skij D. O. (dobrovolskij@gmail.com)
Russian Language Institute of the RAS, Moscow, Russia

Zalizniak Anna A. (anna.zalizniak@gmail.com)
Institute of Linguistics of the RAS; Institute of Informatics Problems of the FRC CSC RAS, Moscow, Russia

The paper outlines the principles of analyzing German and Russian modal constructions. Our first task is to clarify the set of meanings of German modal verbs and the conditions for their implementation. The second task is to describe the means of expressing modal values in Russian that are encountered in parallel corpora as functional equivalents of constructions with German modal verbs. As empirical data we use a representative array of parallel German-Russian texts from the Russian National Corpus (RNC). A supracorpora database of translation correspondences is constructed, in which both the German constructions with modal verbs and their Russian translation equivalents are attributed an annotation of their relevant characteristics. This database, on the one hand, is a valuable linguistic resource that can be used, among other things, to create a new generation of electronic interactive German-Russian and Russian-German dictionaries. On the other hand, the inventory of Russian construction types with (implicit) modal meanings constructed on this database will contribute to the Construction Grammar and confirm the continuity between grammar and lexicon.

Key words: modality, modal verbs, Russian, German, parallel corpus, supracorpora database, bilingual lexicography

1. Постановка задачи

Наше исследование преследует двоякую цель. С одной стороны, это уточнение номенклатуры значений немецких модальных глаголов и условий их реализации, необходимое для дальнейшего развития немецко-русской двуязычной лексикографии. С другой стороны, результаты планируемого исследования позволят построить описание фрагмента грамматики русских конструкций с (потенциально) модальным значением. Как известно, в отличие от немецкого языка, обладающего солидной системой модальных глаголов, в русском языке модальные значения часто выражаются не модальным глаголом, а другими средствами: с помощью сентенциальных наречий *возможно, очевидно, вероятно, наверняка, точно*, предикативов типа *должен, вынужден, можно, нельзя,*

надо, вводных словосочетаний *может быть, должно быть, казалось бы, как представляется*, форм наклонения (условного, повелительного), а также множеством других средств и конструкций «малого синтаксиса», которые до сих пор не выявлены (ср. Падучева 2016). Планируемая база данных аннотированных межъязыковых соответствий послужит инструментом решения обеих этих задач.

Литература, посвященная немецким модальным глаголам, практически безгранична. Из монографий и известных сборников статей последних десятилетий можно назвать Diewald 1999; Fabricius-Hansen et al. 2002; Letnes, Vater 2008; Baumann 2017. Модальные глаголы хорошо описаны в грамматиках; ср. Eisenberg 1999. Что касается сопоставительных исследований (особенно работ, основанных на анализе корпусных данных), то их крайне мало. Сопоставление с русским языком проводилось в книге Weidner 1986, написанной еще в докорпусную эпоху. Из корпусно-ориентированных работ можно назвать исследование Milan 2001 по сопоставлению данных немецкого и итальянского языков. В работе Banasova 2013 немецкие модальные глаголы сопоставляются с их словацкими переводами. Некоторые результаты этого исследования представляются важными для лексикографического представления семантики модальных глаголов, в частности в двуязычном словаре.

Однако задача описания семантики немецких модальных глаголов, опирающегося на отражающие реальное употребление корпусные данные, не решена даже в первом приближении. Учебники и грамматики (даже академические) весьма неполно описывают значения модальных глаголов, не говоря уже о специфических конструкциях, в которые они входят, а имеющиеся параллельные корпуса не позволяют пользователю определить, в каком из возможных значений употреблен тот или иной модальный глагол в каждом из найденных контекстов. Двуязычный словарь, в лучшем случае, дает только весьма приблизительную систематизацию центральных значений многоплановой языковой единицы, которые, как правило, плохо согласуются с конкретными контекстными условиями или иллюстрированы довольно простыми примерами, имеющими слабую объяснительную силу.

Работа по созданию надкорпусной базы данных (НБД)[2] межъязыковых соответствий в области модальных конструкций находится пока в начальной фазе, и настоящая статья обрисовывает общие принципы исследования и используемые методы. Задача статьи — показать, насколько предлагаемый аналитический инструмент полезен для решения очерченного в работе круга проблем.

Ключевым принципом представления межъязыковых соответствий в НБД является разметка в форме набора значений релевантных признаков — морфосинтаксических и семантических (включающих отсылку к значению модального глагола в переводном словаре, типу речевого акта и пр.), а также дистинктивные маркеры контекста. В настоящее время точное и согласованное аннотирование данных по таким признакам может быть обеспечено только в ходе

[2] Принципы построения надкорпусных баз данных изложены в Кружков 2015; см. также Зализняк 2016.

разметки, осуществляемой вручную командой лингвистов-экспертов, хорошо владеющих немецким и русским языком и знакомых с литературой по модальности, а также с методами корпусных исследований.

Исследование ориентируется на идеи и методы теории Грамматики конструкций. Если в классической структуралистской теории принимался постулат, что основу языковой системы составляет грамматика, т. е. совокупность правил порождения сложных единиц из простых по определенным схемам, и лексикону отводилась роль заполнения слотов этих схем, то Грамматика конструкций утверждает, что элементы, из которых строится высказывание, чувствительны к своему окружению и при этом влияют на сам способ построения синтаксических групп. Иными словами, план содержания единиц лексикона зависит от синтаксиса, а синтаксис зависит от выбора лексических единиц. Применительно к модальности это означает, что часто невозможно в каждом конкретном случае выделить лексическую единицу, несущую модальное значение. Часто модальность передается самой синтаксической структурой. Например, такому простому немецкому предложению, как *Was soll ich machen?* в естественном дискурсе соответствует скорее русское *Что мне делать?*, чем буквальное соответствие *Что я должен делать?* (которое, кстати, имеет не совсем то же самое значение, а передает в большей степени идею обязанности). При этом модальное значение в предложении типа *Что мне делать?* передается целиком инфинитивной конструкцией с дативным субъектом в вопросительной форме (что уже отмечалось в специальной литературе, посвященной способам выражения модальности в русском языке).

Существующий в настоящий момент пилотный вариант НБД включает более 600 аннотированных межъязыковых соответствий. Планируемый объем базы данных, на основании которого можно будет делать обобщения, — несколько тысяч записей. В ходе работы возможно будет добавлять новые классификационные признаки, а также изменять определение любого признака, не меняя его названия. В дальнейшем база данных может, естественно, пополняться. В перспективе предполагается также включение направления перевода «русский — немецкий» (т. е. будут исследованы те случаи, когда в переводе на немецкий язык появляется конструкция с модальным глаголом — независимо от того, какой способ выражения модального значения был использован в русском оригинале).

2. Принципы морфосинтаксического аннотирования

Существующие контрастивные лексикографические описания (даже самые современные) учитывают далеко не все *модели перевода*[3] немецких модальных глаголов на русский язык, которые реально используются (о чем можно

[3] «Модель перевода» (translation pattern) — принятый в контрастивных корпусных исследованиях термин (ср. McEnery et al. 2006; Kitamura, Matsumoto 2005). В работе Добровольский и др. 2005 был предложен термин «функционально эквивалентный фрагмент» (далее — ФЭФ).

судить в частности по параллельным текстам НКРЯ), а также обычно недостаточно учитывают влияние контекста на выбор адекватного перевода. Иначе говоря, в параллельных текстах НКРЯ имплицитно присутствуют модели перевода, которые не учтены в описаниях и не включены в словари: установление модели перевода не входит в функциональность параллельных корпусов, кроме того, это выходит за рамки стандартного лексикографического формата[4]. Применяемая в данном проекте методология построения аннотированных переводных соответствий обеспечивает возможность эксплицировать эти модели, что позволит объединить преимущества корпусного подхода и лексикографического описания в рамках двуязычного электронного словаря нового типа.

Так, в примере (1) из немецко-русского параллельного корпуса реализуется значение 3 глагола *sollen*, описываемое в находящемся в работе «Немецко-русском словаре: актуальная лексика» (см. НРС, в печати) следующим образом: «только в формах praet conj и pqp conj *следовало (бы)* (по мнению говорящего); <…> с отрицанием выражает нежелательность: *нежелательно; нельзя*».

(1) was ihm *zur Zufriedenheit* ***hätte*** *gereichen* ***sollen***, flößte ihm Grauen ein. [Friedrich Dürrenmatt. Der Verdacht (1953)] — ***вместо*** *удовлетворения* он испытывал ужас. [Фридрих Дюрренматт. Подозрение (Н. Савинков, 1990)]

Здесь в качестве единицы перевода реально берется не отдельный модальный глагол и даже не его непосредственное окружение (зависимый инфинитив с заполненными валентностями), а целая клауза *was ihm zur Zufriedenheit hätte gereichen sollen*. Эта клауза переводится предложной группой *вместо удовлетворения*. Модальный смысл, выражаемый в немецком языке плюсквамперфектной формой конъюнктива модального глагола *sollen* [*hätte* + Inf I + *sollen*], сконцентрирован в русском предлоге *вместо*. При менее идиоматичном и в какой-то степени искусственном переводе это предложение могло бы выглядеть как *То, что следовало бы <должно было бы> дать ему удовлетворение, наполнило его ужасом*. Переводчик, однако, воспользовался иной конструкцией, не содержащей модального слова, но при этом достаточно полно отражающей смысл исходной немецкой фразы. Этот способ передачи модального значения, не учтен ни в одном описании русско-немецких соответсвий в области способов выражения модальности.

На этом примере видно, как обращение к параллельным текстам НКРЯ позволяет выявить реально представленные в узусе модели перевода модального глагола в рамках уже зафиксированного в словаре значения. Единицей создаваемой надкорпусной базы данных является *моноэквиваленция*[5] т. е. двухместный кортеж <немецкий оригинал, содержащий модальный глагол; ФЭФ немецкой конструкции с модальным глаголом в русском переводе>, где каждый из двух элементов снабжен формализованной *аннотацией* (= набором тегов, представляющих собой значения классификационных признаков). Формат

[4] Об использовании параллельных корпусов в сопоставительной лексикологии и в немецко-русской двуязычной лексикографии см. Добровольский 2015.

[5] Термин введен в Loiseau et al. 2013.

планируемой базы данных таких соотвествий позволит осуществлять поиск как по лексемам, так и по всем классификационным признакам, используемым в аннотациях, для обоиз языков.

Так, немецкая фраза в примере (1) будет снабжена следующей аннотацией:

sollen:
<+Inf I> — управляет глаголом в форме инфинитива I;
<3sg> — 3 лицо, единственное число;
<Konj II> — Konjunktiv II;
<Pqp> — Plusquamperfekt;
<sollen-3> — соответствует 3-му значению словарной статьи *sollen* в НРС.

В примере (2) немецкая конструкция [*müssen* + Inf] переведена не содержащей модального глагола градационной конструкцией [*не столь… чтобы* + Inf].

(2) Er war, als er heranwuchs, nicht besonders groß, nicht stark, zwar häßlich, aber nicht so extrem häßlich, dass man vor ihm ***hätte*** *erschrecken* ***müssen***. [Patrick Süskind. Das Parfum: Die Geschichte eines Mörders (1985)] — Подростком он был не слишком высок, не слишком силен, пусть уродлив, но ***не столь*** исключительно уродлив, ***чтобы*** *пугаться* при виде его. [Патрик Зюскинд. Парфюмер: История одного убийцы (Э. Венгерова, 1992)][6]

Сходная ситуация имеет место в примере (3). Форма Konjunktiv II модального глагола *müßte* на первый взгляд остается без перевода. Однако за модальность явно отвечает русская единица *как будто*. Важно, однако, что здесь нельзя говорить об эквивалентности форм *müßte* и *как будто*: эквивалентными оказываются конструкции [*müßte geradezu* P] и [*как будто* P]. Выражаемое значение может быть приблизительно описано как ‘почти что можно сказать, что…’.

(3) Übrigens scheint Berta Lunte zu riechen, sie sagt, die Streichhölzer ***müßte*** *geradezu* jemand fressen. [Erich Kästner. Pünktchen und Anton (1931)] — Кстати, Берта, кажется, что-то пронюхала, она говорит, что в доме уходит прорва спичек, ***как будто*** их кто-то жрет. [Эрих Кестнер. Кнопка и Антон (Е. Вильмонт, 2001)]

Пример (4) иллюстрирует способ представления переводного соответствия, используемый в НБД (ср. также пример (5)).

Здесь тег <+Inf I> указывает на сочетание модального глагола с формой Infinitiv I. Тег <Sie> указывает на вежливую форму обращения (ко 2-му лицу). Тег <Praet> показывает, что модальный глагол стоит в форме Präteritum, а тег <Konj II> на форму сослагательного наклонения Konjunktiv II. Тег <sollen-3> означает, что глагол *sollen* в данном предложении употреблен в 3-м значении словарной статьи *sollen* в НРС.

[6] В контекстах из НКРЯ сохранена орфография оригинала. Часто это орфографическая норма до реформ правописания 1996, 2005 и 2006 годов.

В русской части выделение п/ж формы *давай* указывает на то, что модальность в русском переводе передается прежде всего этой формой императива (этот компонент является «строевым»), а выделение курсивом формы *оставим в покое* — на обязательное заполнение валентности формы *давай*, образующей конструкцию аналитического императива. Тег <конструкция «**давай** + 1pl Pfv»> указывает на использование аналитической формы императива совместного действия, включающей зависимый глагол в форме 1 л. мн. ч. сов. вида. Теги <Imperat> и <2sg> указывают на характеристики формы *давай* в составе этой конструкции.

(4)	*Sie* **sollen** diese ungemütliche Sache jetzt *lieber sein lassen*. [Friedrich Dürrenmatt. Der Verdacht (1953)]	**sollen** <+Inf I> <Sie> <Praet> <Konj II> <sollen-3>	— **Давай** *оставим* это дело *в покое*. [Фридрих Дюрренматт. Подозрение (Н. Савинков, 1990)]	<конструкция «**давай** + 1pl Pfv»> <Imperat> <2sg>

Данный пример демонстрирует неочевидное (не фиксируемое словарями) переводное соответствие между немецкой конструкцией [*Sie sollten* + Inf I] и русской аналитической формой императива совместного действия с *давай*.

(5)	Schwester Luise hatte, *wie es die Mutter* **gewollt**, das schöne Kleid angezogen, [Ernst Theodor Amadeus Hoffmann. Nußknacker und Mausekönig (1816)]	**wollen** <– Inf> <3sg> <Pqp> <wollen-1>	Старшая сестра Луиза, *по* **желанию** *матери*, надела [...] нарядное платье, [Эрнст Теодор Амадей Гофман. Щелкунчик и Мышиный король (И. Татаринова, 1937)]	**желание** <абстрактное существительное> <*по* + Dat>

Тег <– Inf I> указывает на то, что модальный глагол *wollen* в данном употреблении (во вводной эллиптической конструкции) не управляет глаголом в инфинитиве. <3sg> — 3 лицо, единственное число; <Pqp> — Plusquamperfekt; <wollen-1> — соответствует 1-му значению словарной статьи *wollen* в НРС.

В аннотации русского перевода указан строевой компонент — абстрактное существительное *желание* и отмечена предложно-падежная форма, в которой оно употреблено.

Отметим, что семантика модального глагола *wollen* 'хотеть' полностью сохраняется в переводе и передается с помощью русского существительного *желание* в составе предложной группы *по желанию матери*, что оказывается более идиоматичным, чем ожидаемая фраза *как хотела мама*, и лучше передает стилистику контекста.

Как видно на этих примерах, немецкая конструкция с модальным глаголом может иметь в качестве переводного эквивалента… предлог (1), конструкции [*не столь… чтобы* + инф.] (2) и [*как будто P*] (3), а также конструкции, ядро которых образует форма императива (4), абстрактного существительного (5).

3. Элементы семантического аннотирования

Семантическое аннотирование включает, прежде всего, приписывание значения данного модального глагола в соответствии с номенклатурой, принятой в НРС. Так, пример (6) с глаголом *sollen* получает тег <sollen-1>, поскольку в этом контексте реализуется значение 1 словарной статьи глагола *sollen* в НРС: «для выражения обязанности — “быть должным” (что-л. делать по чьему-л. указанию, по закону, по правилам и т. п.)».

(6) Verhalten bei einem Stopp der SPS-CPU wählbar Bei der Parametrierung kann festgelegt werden, ob Daten, die an dezentrale Stationen, lokalen Stationen, einem intelligenten Gerät oder einem Stand by-Master übertragen werden, weiter aktualisiert oder gelöscht werden ***sollen***, wenn die SPS-CPU gestoppt ist. — Возможность выбора режима при останове центрального процессора контроллера, в параметрах можно установить, должны ли при останове центрального процессора контроллера продолжать обновляться данные, передаваемые на удаленные, локальные, интеллектуальные или резервную ведущую станцию, или эти данные ***должны*** стираться. [Mitsubishi Europe. Speicherprogrammierbare Steuerungen Bedienungsanleitung [ABBYY LingvoPRO] (2012)]

Такая разметка поможет установить, какое из значений данного модального глагола реализуется в данном контексте и какие способы перевода этого контекста на русский язык оказываются наиболее вероятными.

Все немецкие модальные глаголы имеют два типа прочтения (ср. Vater 2001: 81): 1) комментарий к истинности содержания пропозиции, так называемое субъективное или эпистемическое прочтение; 2) как часть содержания высказывания, так называемое объективное или деонтическое прочтение. Ср. пример (7) (заметим, что эта амбивалентность присутствует не только в немецкой фразе (7), но и в ее русском переводе):

(7) Auch er ***müsse*** einen Schlüssel zu dem Pförtchen besitzen, sagte sie ihm. [Gottfried Keller. Der grüne Heinrich (zweite Fassung) (1879–1880)] У него тоже ***должен*** быть ключ к калиточке, сказала она. [Готфрид Келлер. Зеленый Генрих (Ю. Афонькин, Г. Снимщикова, Д. Горфинкель, Н. Бутова, 1972)]

Чтобы понять, какой тип прочтения имеется в виду, необходимо обращаться к более широкому контексту, который может быть предоставлен в аутентичных текстах с их профессионально выполненными русскими переводами — ср. пример (8), из которого ясно, что имеется в виду эпистемическое прочтение:

(8) Auch er ***müsse*** einen Schlüssel zu dem Pförtchen besitzen, sagte sie ihm; *er holte einen Kasten mit allerlei alten Schlüsseln herbei und fand mit ihrer Hilfe richtig denjenigen heraus, welcher in das Schloß paßte*. [Gottfried Keller. Der grüne Heinrich (zweite Fassung) (1879–1880)] У него тоже ***должен*** быть ключ к калиточке, сказала она. Альбертус *сейчас же притащил ящик со всякими старыми ключами и среди них, с ее помощью, нашел тот, что подходил к замку.* [Готфрид Келлер. Зеленый Генрих (Ю. Афонькин, Г. Снимщикова, Д. Горфинкель, Н. Бутова, 1972)]

Функциональность создаваемой базы данных позволяет учитывать, в том числе, наличие контекста, разрешающего подобную неоднозначность.

Соответственно, пример (7) — как и его расширенный вариант (8) — получает тег <muessen-4>, отсылающий к значению этого глагола в НРС «в значении предположения, б. ч. основанного на фактах; тж. в *praet conj* выражает предположение с меньшей степенью уверенности».

Существенность признака «степень уверенности» предположения, который также используется в разметке, может быть продемонстрирована на примере эпистемических значений глаголов *dürfen* и *können*.

Глагол *dürfen* в одном из значений употребляется большей частью в формах претеритума конъюнктива (Praet Konj) в сочетании с инфинитивом II (Inf II), а соответствующие конструкции переводятся с использованием русских модальных выражений *по всей видимости*, *вероятно*; *пожалуй*. Эти конструкции употребляются для выражения предположения с высокой степенью уверенности: *sie* ***dürften*** *schon schlafen — они,* ***вероятно****, уже спят*; *das* ***dürfte*** *genügen — этого,* ***пожалуй****, достаточно*; *das Geld* ***dürfte*** *schon überwiesen worden sein —* ***скорее всего****, деньги были уже переведены на счёт*.

Глагол *können* употребляется в одном из значений в аналогичных формах, но выражает предположение с равными шансами его осуществления или неосуществления. Ср.: *der Arzt* ***kann*** *in einer Stunde hier sein —* (***думаю, что***) *врач придёт не раньше, чем через час*; *ich weiß nicht genau, aber dieser Einfall* ***könnte*** *von unserem Chef stammen — я точно не знаю, но* ***похоже****, что это идея нашего начальника*; *er kann sein neues Auto schon gekauft haben —* ***может быть*** (***возможно*** (***что***)) *он уже купил новую машину*.

В аннотировании отражаются также такие признаки как «вопросительное предложение», «отрицание», «перформативность». Ср. примеры (9) и (10).

В вопросительном предложении конструкция с модальным глаголом часто переводится на русский язык принципиально иным способом, чем в утвердительном — особенно, если это вопрос с отрицанием. В русском переводе в таких случаях часто используется конструкция [*не… ли*], как в (9), или [*разве не*].

(9) War nicht sein Vater längst gestorben, allein, ohne seinen Sohn wiedergesehen zu haben? ***Musste*** *er* selbst ***nicht*** dies selbe Schicksal *erwarten*? [Hermann Hesse. Siddhartha (1922)] — Не заставил ли он своего отца страдать столько же, сколько он страдал? теперь из-за своего собственного сына? ***Не*** *ожидает* ***ли и*** *его*, Сиддхартху, такая же участь? [Герман Гессе. Сиддхартха (Б. Д. Прозоровская, 1990)]

Пример (10) показывает, что в определенных случаях функция модального глагола *müssen* сводится к перформативности: *ich muss gestehen/zugeben*.

(10) ***Ich muss*** *dir* ***gestehen***, Lieber: ich unterscheide zwischen Gedanken und Worten nicht sehr. [Hermann Hesse. Siddhartha (1922)] — *Только* ***признаюсь*** *тебе*, мой милый: я не вижу большого различия между мыслями и словами. [Герман Гессе. Сиддхартха (Б. Д. Прозоровская, 1990)]

В русском языке в таких контекстах используется будущее время (*призна́юсь*, как в (10)), хотя в принципе возможен и более близкий к дословному перевод *должен признаться*. В подобных случаях в переводе может появлятся, казалось бы, противоположный по семантике глагол *мочь*; ср. *ich muss sagen, dass… — могу сказать, что…*

В семантическом аннотировании конструкций с модальным глаголом предполагается также использовать постепенно утверждающийся в специальной литературе (ср. Divjak et al. 2015) подход, согласно которому языковые выражения модальности должны классифицироваться не по типам (таким, как онтологическая, эпистемическая, деонтическая, алетическая, волитивная модальность), а по конкретным модальным значениям (таким как «возможность») и по соответствующим речевым актам (таким как «разрешение» или «запрет») — ср., например, значение «предположения» в примерах (7)–(9).

4. Заключение

Создаваемая НБД послужит инструментом для решения одновременно двух задач. С одной стороны, такая база данных предоставит в наше распоряжение детальную, многоаспектную и подкрепленную аутентичными контекстами информацию о реальном употреблении немецких конструкций с модальными глаголами и об их возможных переводных соответствиях в русском языке, которая может быть использована при разработке концепции двуязычного электронного словаря нового поколения, включающем интерактивную отсылку к параллельному корпусу. С другой стороны, систематизация моделей перевода немецких конструкций с модальными глаголами позволит существенно расширить наши представления о возможностях передачи модальных значений средствами русского языка, а также произвести выявление и инвернтаризацию типов русских модальных конструкций, многие из которых еще не были описаны.

К настоящему моменту разработана общая архитектура надкорпусной базы данных, выработана методика аннотирования немецких конструкций с модальными глаголами и их соответствий в русском языке и сформирован основной набор релевантных классификационных признаков (тегов), по которым осуществляется морфосинтаксическое и семантическое аннотирование. Полученные на данном этапе предварительные результаты подтверждают эффективность предлагаемых методов.

Литература

1. *Добровольский Д. О.* Корпус параллельных текстов и сопоставительная лексикология // Труды института русского языка им. В. В. Виноградова. Вып. 6. М., 2015. С. 411–446.
2. *Добровольский Д. О., Кретов А. А., Шаров С. А.* Корпус параллельных текстов: архитектура и возможности использования // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М., 2005. С. 263–296.

3. *Зализняк Анна А.* База данных межъязыковых эквиваленций как инструмент лингвистического анализа // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог 2016. М., 2016. С. 763–775.
4. *Кружков М. Г.* Информационные ресурсы контрастивных лингвистических исследований: электронные корпуса текстов // Системы и средства информатики, 2015. Т. 25. № 2. С. 140–159.
5. *НРС* = Немецко-русский словарь: актуальная лексика / Под ред. Д. О. Добровольского. М.: Лексрус, в печати.
6. *Падучева Е. В.* Модальность // Материалы к Корпусной грамматике русского языка. Глагол. Часть I. СПб.: Нестор-История, 2016. С. 19–94.
7. *Banasova M.* Deutsche Modalverben und ihre Äquivalente im Slowakischen. Berlin: Logos Verlag, 2013.
8. *Baumann C.* Bedeutung und Gebrauch der deutschen Modalverben: Lesarten und besondere Verwendungsweisen zwischen lexikalischer Einheit und kontextueller Vielheit der Modalverbbedeutung. Berlin: Walter de Gruyter, 2017.
9. *Divjak, D., Szymor N., Socha-Michalik A.* Less is more: possibility and necessity as centres of gravity in a usage-based classification of core modals in Polish // Russian Linguistics 39(3). 2015. Pp. 327–349.
10. *Diewald G.* Die Modalverben im Deutschen: Grammatikalisierung und Polyfunktionalität. (= Germanistische Linguistik. 208). Tübingen: Niemeyer, 1999.
11. *Eisenberg P.* Grundriß der deutschen Grammatik. Band 2: Der Satz. Stuttgart: Metzler, 1999.
12. *Fabricius-Hansen C., Oddleif L., Letnes O.* (Hrsg.). Modus, Modalverben, Modalpartikeln. Trier: WVT, 2002.
13. *Kitamura M., Matsumoto Y.* Practical Translation Pattern Acquisition from Combined Language Resources // Natural Language Processing — IJCNLP 2004. Berlin etc.: Springer, 2005. Pp. 244–253.
14. *Letnes O., Vater H.* (Hrsg.). Modalität und Grammatikalisierung = Modality and grammaticalization. Trier: WVT, 2008.
15. *Loiseau S., Sitchinava D. V., Zalizniak Anna A., Zatsman I. M.* Information technologies for creating the database of equivalent verbal forms in the Russian-French multivariant parallel corpus // Информатика и ее применения, 2013. Том 7, вып. 2. С. 100–109.
16. *McEnery T., Xiao R., Tono Y.* Corpus-based Language Studies: An Advanced Resource Book. London, NY: Routledge, 2006.
17. *Milan C.* Modalverben und Modalität: Eine kontrastive Untersuchung Deutsch-Italienisch. Berlin: Walter de Gruyter, 2001.
18. *Vater H.* (2001) *Sollen* und *wollen* — zwei ungleiche Brüder. In: Vater H., Letnes O. (Hrsg.) Modalität und mehr / Modality and more. Trier: WVT, S. 81–100.
19. *Weidner A.* Die russischen Übersetzungsäquivalente der deutschen Modalverben: Versuch einer logisch-semantischen Charakterisierung. München: Otto Sagner, 1986.

References

1. *Banasova M.* (2013), German Modal Verbs and Their Equivalents in Slovak [Deutsche Modalverben und ihre Äquivalente im Slowakischen]. Logos Verlag, Berlin.
2. *Baumann C.* (2017), Meaning and Use of German Modal Verbs: Readings and Special Uses between Lexical Unity and Contextual Plurality of Modal Verb Meanings [Bedeutung und Gebrauch der deutschen Modalverben: Lesarten und besondere Verwendungsweisen zwischen lexikalischer Einheit und kontextueller Vielheit der Modalverbbedeutung], Walter de Gruyter, Berlin.
3. *Divjak, D., Szymor N., Socha-Michalik A.* (2015), Less is more: possibility and necessity as centres of gravity in a usage-based classification of core modals in Polish, Russian Linguistics 39(3), pp. 327–349.
4. *Diewald G.* (1999), The Modal Verbs in German: Grammaticalization and Polyfunctionality. [Die Modalverben im Deutschen: Grammatikalisierung und Polyfunktionalität]. Niemeyer, Tübingen.
5. *Dobrovol'skij D. O.* (2015), The corpus of parallel texts and contrastive lexicology [Korpus parallel'nykh tekstov i sopostavitel'naya leksikologiya], Proceedings of the V. V. Vinogradov Russian Language Institute [Trudy instituta russkogo yazyka im. V. V. Vinogradova], Issue 6, Moscow, pp. 411–446.
6. *Dobrovol'skij D. O., Kretov A. A., Sharov S. A.* (2005), Corpus of parallel texts: architecture and use possibilities [Korpus parallel'nykh tekstov: arkhitektura i vozmozhnosti ispol'zovaniya], The National Corpus of the Russian language: 2003–2005. Results and prospects [Natsional'nyy korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy], Moscow, pp. 263–296.
7. *Eisenberg P.* (1999), Foundations of the German Grammar [Grundriß der deutschen Grammatik], Volume 2: The Sentence [Band 2: Der Satz], Metzler, Stuttgart.
8. *Fabricius-Hansen C., Oddleif L., Letnes O.* (ed.) (2002). Mode, Modal Verbs, Modal Particles [Modus, Modalverben, Modalpartikeln], WVT, Trier.
9. *Kitamura M., Matsumoto Y.* (2005), Practical translation pattern acquisition from combined language resources, Natural Language Processing — IJCNLP 2004, Springer, Berlin etc., pp. 244–253.
10. *Kruzhkov M. G.* (2015), Information resources of contrastive linguistic research: Electronic text cases [Informatsionnyye resursy kontrastivnykh lingvisticheskikh issledovaniy: elektronnyye korpusa tekstov], Systems and Means of Informatics, [Sistemy i sredstva informatiki], Vol. 25, No. 2. pp. 140–159.
11. *Letnes O., Vather H.* (ed.) (2008). Modality and Grammaticalization [Modalität und Grammatikalisierung], WVT, Trier.
12. *Loiseau S., Sitchinava D. V., Zalizniak Anna A., Zatsman I. M.* (2013), Information technologies for creating the database of equivalent verbal forms in the Russian-French multivariant parallel corpus, Computer science and its applications [Informatika i ee primeneniya], Vol. 7, No. 2, pp. 100–109.
13. *McEnery T., Xiao R., Tono Y.* (2006), Corpus-based Language Studies: An Advanced Resource Book. Routledge, London, NY.

14. *Milan C.* (2001), Modal Verbs and Modality: A Contrastive Examination German-Italian [Modalverben und Modalität: Eine kontrastive Untersuchung Deutsch-Italienisch], Walter de Gruyter, Berlin.
15. *HPC* = German-Russian Dictionary: Actual Lexicon [Nemetsko-russkiy slovar': aktual'naya leksika] (in print), ed. by D. O. Dobrovol'skij, Lexrus, Moscow.
16. *Paducheva E. V.* (2016), Modality [Modal'nost'], Materials to the Corpus Grammar of Russian [Materialy k Korpusnoy grammatike russkogo yazyka], Part I. Verb [Glagol], Nestor-Istoriya, St. Petersburg, pp. 19–94.
17. *Vater H.* (2001), *Should* and *want* — two unequal brothers [*Sollen* und *wollen* — zwei ungleiche Brüder], Modality and more [Modalität und mehr], ed. by Vater H., Letnes O., WVT, Trier, pp. 81–100.
18. *Weidner A.* (1986), The Russian translation equivalents of the German modal verbs: Attempt of a logical-semantic characterization [Die russischen Übersetzungsäquivalente der deutschen Modalverben: Versuch einer logisch-semantischen Charakterisierung], Otto Sagner, Munich.
19. *Zalizniak Anna A.* (2016), Database of cross-linguistic equivalents as a tool of linguistic analysis [Baza dannykh mezh"yazykovykh ekvivalentsiy kak instrument lingvisticheskogo analiza], Computer linguistics and intellectual technologies. Proceedings of the international conference "Dialog 2016" [Komp'yuternaya lingvistika i intellektual'nyye tekhnologii. Po materialam mezhdunarodnoy konferentsii "Dialog 2016"]. Moscow, pp. 763–775.

ДИСКУРСИВНЫЙ МАРКЕР *ТИПА* ПО ДАННЫМ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА: ПРОИСХОЖДЕНИЕ, СЕМАНТИКА И ПРАГМАТИКА

Егорова М. А. (drajenka@gmail.com)
РГГУ, Москва, Россия

DISCOURSE MARKER *TIPA* ACCORDING TO THE DATA OF RUSSIAN NATIONAL CORPUS: ITS ORIGIN, SEMANTICS AND PRAGMATICS

Egorova M. A. (drajenka@gmail.com)
RSUH, Moscow, Russia

Discourse marker *tipa* became widespread in colloquial Russian in the decade 1990s–2000s. However, until recently, it has gained little attention. In this paper we use the data from the Russian National Corpus and we aim to accomplish the following goals: 1) to highlight the origin of the discourse marker *tipa* from the noun *tip* 'type', 2) to describe the semantics of the discourse marker *tipa* as well as that of the partly grammaticalized element *tipa* as part of parametric constructions. We base our approach mainly on the results achieved by Susanne Fleischman and Marina Yaguello.

Key words: discourse marker, "xeno" marker, hedge, semantics

1. Вступление

Дискурсивная частица *типа* используется в современном русском языке главным образом в неформальной спонтанной речи, напр.[1]:

(1) Может, это отсвет, конечно, от крыши, может, это закат был какой-нибудь / или прочее, / но, тем не менее, / я утверждаю, что это женщина / в розовых тонах. // Вот она, / *типа*, / бежит. / «*Типа*» — это значит «как будто», / это сленг современный. / Как видите, даже / вот в такую / сравнительно элитную среду, к которой я принадлежу / (да чего там!), / тоже проникла эта / ядовитость, // да, гадость нашего времени. [Описание изображений // Из коллекции А. А. Степихова, 2003]

Пример (1) интересен тем, что высказывание становится в нём объектом рефлексии говорящего, который отметил важные свойства интересующей нас частицы — её аппроксимативную семантику, новизну и субстандартный характер.

Корпусные данные показывают, что, хотя частица *типа* спорадически встречается в устной речи уже в 80-е годы прошлого века, широкое распространение она получает начиная с середины 1990-х годов. Так, в период с 1987 по 1996 год её общая частота в единицах ipm в подкорпусе устной непубличной речи Национального Корпуса Русского Языка равна 48,8, а в период с 1997 по 2006 год — 505, т. е. на порядок больше[2].

2. Постановка задачи

Несмотря на то, что частица *типа* уже привлекала внимание лингвистов[3], она, насколько нам известно, не становилась предметом специального исследования, и одной из наших задач является восполнить этот пробел.

Новизна данной языковой единицы делает особенно наглядной её связь с существительным *тип* и частично декатегоризованным *типа* [4]в составе параметрических конструкций вида ‘х типа у’ (*звёзды типа Солнца*, *сверхновые типа один*)[5]. У нас есть все основания полагать, что дискурсивный маркер *типа*

---

[1] Этот пример, как и большинство русских примеров в данной статье, взят из Национального корпуса русского языка (НКРЯ). В тех случаях, когда пример взят из другого источника, на это будет указано специально.

[2] См. [Подлесская, Егорова 2017].

[3] См., напр., [Плунгян 2008], [Левонтина 2010], [Подлесская, Стародубцева 2013].

[4] В русской грамматической традиции *типа* в таких конструкциях рассматривается как предлог с аппроксимативным значением [Русская грамматика 1980, Том I, с. 706] или как неизменяемая единица, по функциям близкая к предлогу, — *предложное образование* [Русская грамматика 1980, Том II, с. 66].

[5] См. [Подлесская, Егорова 2017].

возник из *muna* в составе параметрических конструкций в результате дальнейшей декатегоризации и прагматизации последнего[6].

Таким образом, *muna* входит в один ряд с дискурсивными маркерами, восходящими к словам со значением подобия, таким как хорошо изученный английский аппроксиматор *like*, и, в частности, к существительным со значением ‘род, вид, тип, сорт’, ср. англ. *kind of*, *sort of* и фр. *genre*, которые с лёгкостью развивают симилятивную семантику[7]. Не имея возможности приводить здесь обширную библиографию, упомянем статью [Fleischman, Yaguello 2004], в которой подробно рассматриваются употребления (‘функции’, *functions*) таких маркеров; там же читатель может найти основную литературу[8].

Все перечисленные в [Fleischman, Yaguello 2004] ‘функции’ дискурсивных маркеров подразделяются на три основные группы: ‘пропозициональные’ (*Propositional*), ‘текстуальные’ (*Textual*) и ‘экспрессивные’ (*Expressive*).

В первую группу входят употребления, ориентированные на объективную оценку действительности: выражение сравнения (*Expression of Comparison*), ср. (2), сходства / подобия (*Similarity*), ср. (3)[9], приблизительности (*Approximation*), ср. (4) и, наконец, так называемый парадигматический пример (*Paradigmatic Example*) — типичный частный случай, который даёт адекватное представление о конкретном объекте /явлении или целом классе объектов / явлений, ср. примеры (5) и (6), соответственно.

(2) She sings *like* a bird.
«Она поёт, *как* птица»

(3) My love is *like* a rose.
«Моя любовь подобна розе»

(4) *Французский*:
Il saute *genre* 1m 30 sans peine.
«Он без усилий прыгает метра на полтора»

(5) J’ai entendu un bruit de camion *genre* poubelle.
«Я услышал(а) шум грузовика, *muna* мусоровоза»

(6) My mother would always say stuff *like*: “Now, don’t go out without your coat!”
«Моя мама вечно говорила вещи *muna*: „Только не выходи на улицу без пальто!“»

Во вторую группу входят фокусные употребления (*Focus*), причём под фокусом понимается наиболее значимая и /или новая информация в предложении. Здесь авторы следуют работе [Underhill 1988], где фокус лаконично

---

[6] Аналогичный процесс — грамматикализация существительных *type*, *kind*, *sort*, в ходе которой последние проходят через стадию частично декатегоризованных единиц в составе т. н. *SKT constructions* — описан в статье [Dehé, Stathi 2016].

[7] См., напр., [Heine, Kuteva 2002] с. 274.

[8] См. также [Fleischman 1998] и [Yaguello 1998].

[9] Пример 3 взят из [Heine, Kuteva 2002], с. 274, со ссылкой на [Fleischman 1999].

определяется как “the point of the sentence”. Примерами фокусных употреблений могут служить (7–8), где основной функцией дискурсивной частицы является маркировать фокус в том смысле, как он понимается выше.

(7) Her parents *like* bought a farm in North Carolina.
«Её родители *типа* купили ферму в Северной Каролине»

(8) *Финский*:
Se kurssi on *niinku* todella vaikee mulle, mä en *niinku* ollenkaan tajuu mistä siellä *niinku* puhutaan.
«Этот курс *типа* действительно сложный для меня, я *типа* вообще не понимаю, о чём они там *типа* говорят»

В третью группу входят употребления, ориентированные в большей степени на передачу отношения говорящего к сказанному: к ним относится функция аппроксиматора (*Hedge*) и так называемый *Interpretive Quotative* — аппроксимативное цитирование[10]. Аппроксиматор выражает оценку избранной номинации или высказывания в целом как не вполне точных, подходящих или достоверных[11]. По удачному выражению С. Флейшман, маркер-аппроксиматор «снимает с говорящего ответственность», снижая достоверность сообщения или смягчая категоричность высказывания[12]. Так, в (9) и (10) маркеры *like* и *genre* — каждый из которых можно перевести как *типа* — смягчают высказывание, кажущееся говорящему слишком смелым:

(9) A.: Come over any time. B: *Like* . . . right now?
«A: Приходи в любое время. B.: *Типа*… прямо сейчас?»

(10) *Французский*:
Je me demandais si tu pourrais *genre* me donner un coup de main.
«Я думал(а), не можешь ли ты *типа* мне помочь»

Аппроксимативное цитирование, помимо указания на неточность (11), нередко включает субъективную оценку чужих слов, интерпретацию мыслей или намерений другого и поэтому подходит для оформления таких высказываний, которые в действительности никогда не звучали, например, внутренней речи (12) или вымышленного высказывания, обобщающего множество частных случаев (13).

(11) And I’m *like*: “What the hell’s going on here?”
«И я *такой / такая*: „Что, чёрт возьми, происходит?“»

---

[10] После выхода статьи [Арутюнова 2000] в русской традиции за маркерами в функции оформителей чужой речи (или собственной речи, воспринимаемой как “отчуждённая”) закрепилось название ‘ксенопоказатели’, или ‘ксеномеркеры’.

[11] Такое узкое понимание термина *hedge* не является общепринятым.

[12] [Fleischman 1998] с. 38. В [Fleischman, Yaguello 2004] функция маркера-аппроксиматора (*Hedge*) определяется как “signaling that whatever is in the markers’ scope should not be taken too literally”.

(12) And I’m *like* OK, how am I gonna get her “chief complaint” out of her?
«И я *типа* о’кей, как же мне выяснить у неё, на что она жалуется?»

(13) Les étrangers, ils nous englobent dans un tout, *genre*: “on est pas surpris, sa se passe toujours comme ça en France”.
«Иностранцы считают нас всех одинаковыми, *типа*: „неудивительно, во Франции всегда так“»

Предложенная классификация нуждается в двух существенных уточнениях.

Во-первых, три перечисленные функции не являются взаимоисключающими[13]. Продемонстрируем это на примере количественной аппроксимации, которую принято считать «ядром аппроксимативной зоны»[14].

Рассмотрим английский пример (14)[15], французский пример (15), уже обсуждавшийся выше под номером (4)[16], и русский пример (16).

(14) Greg is *like* seven or eight feet tall.
«Грэг футов семь-восемь ростом»

(15) Il saute *genre* 1m 30 sans peine.
«Он без усилий прыгает метра на полтора»

(16) [Лиза] Это Тышлер. Был.
[Вика] В смысле?
[Лиза] Художник такой.
[Вика] Дорогой?
[Лиза] Очень. От дедушки остался.
[Вика] Слушай, ужас-то какой. Чё, прям совсем дорогой? *Типа* миллион?
[Авдотья Смирнова, Анна Пармас. Кококо, к/ф, 2012]

Количество — типичный пример информации, которую говорящий в ряде случаев может сообщить лишь приблизительно. При этом он может считать, что выбранная им номинация является вполне удачной. Например, в (15) говорящий вполне уверен, что субъект высказывания без труда способен прыгнуть на расстояние, примерно равное полутора метрам. А в (16) героиня фильма, испортив картину, пытается выяснить её примерную стоимость, называя число, которое при дальнейшем обсуждении будет скорректировано. Поэтому мы полагаем, что в этих случаях на первый план выходит соотношение высказывания с действительностью, а не субъективная оценка собственных слов,

---

[13] Авторы делают и более сильное утверждение, а именно: что дискурсивные маркеры приобретают названные функции в определённом порядке — вначале ‘пропозициональные’, вслед за ними ‘текстуальные’ и лишь затем ‘экспрессивные’, причём этому утверждению придаётся статус универсалии. Но именно в силу того, что функции не исключают друг друга, данное утверждение с трудом поддаётся проверке.

[14] [Подлесская, Стародубцева 2013] с. 30.

[15] Пример 14 взят из [Fleischman, Yaguello 2004] со ссылкой на [Underhill 1988].

[16] Пример 16 отсутствует в НКРЯ.

и, следовательно, в примерах (14–16) дискурсивная частица выражает пропозициональное значение приблизительности (*Approximation*).

Такое значение приписывается маркерам в (14–15) в [Fleischman, Yaguello 2004]. Но, поскольку, помимо приблизительности, дискурсивные частицы *like / genre* вводят новую для слушателя информацию, о росте Грега в (14) и о длине прыжка в (15), эти же примеры рассматриваются в [Fleischman 1998] и [Fleischman, Yaguello 2004] и в контексте обсуждения фокусных употреблений. А если бы мы имели основания считать, что в (14) говорящий испытывает значительные сомнения относительно точного роста Грега, *like* мог бы рассматриваться и как маркер-аппроксиматор. Сказанное верно и по отношению к 16: *типа*, помимо выражения приблизительности, может придавать вопросу оттенок неуверенности или несерьёзности.

Таким образом, на данном этапе у нас нет строгой формальной процедуры для разграничения значений, как нет её и в обсуждаемых выше работах, и нам во многом приходится полагаться на интуицию. Несмотря на это, предложенная классификация представляется нам удобной отправной точкой для того, чтобы систематизировать наши наблюдения. Вначале мы кратко остановимся на *типа* в составе параметрических конструкций[17], а затем перейдём к набору значений омонимичной дискурсивной частицы.

3. *Типа* в составе параметрических конструкций вида ‘x типа y’

В конструкциях вида ‘x типа y’ ‘X’ называет класс (множество) сущностей, внутри которого выделяется подкласс (подмножество), специфицируемое при помощи ‘Y’. ‘Y’ может прямо называть выделяемый подкласс, тем самым чётко очерчивая его границы (17):

(17) Завтрашний день по просьбе трудящихся будет отмечен *кремовым тортом типа «Наполеон»*. [Виктор Макаров и др. Берегите женщин, к/ф, 1981]

(18) Уже накоплено столько данных в самых-самых разных науках / и генетике / и антропологии / и этологии / и когнитивной науке / и многого-многого другого / что теперь уже невозможно построить *гипотезу типа сказки*. [Светлана Бурлак. О неизбежности происхождения человеческого языка. Лекции Полит.ру, 2008]

(19) Он показал / что да / действительно / *органические молекулы типа там и липидов / и аминокислот могут собираться в капли*. [Возникновение биологической информации. Программа «Гордон» (НТВ), 2003]

(20) Интеллигент / трезво оценивающий момент / или *слепой фанатик… типа Кальтенбруннера*? [Татьяна Лиознова, Юлиан Семенов. Семнадцать мгновений весны, к/ф, 1973]

[17] Семантике конструкций со словами *типа*, *вроде*, *наподобие* посвящена статья [Savchenko 2015]. Наше описание преследует иные цели и будет по необходимости более кратким.

В других случаях ‘Y’ не называет подмножество целиком, но позволяет составить представление о нём, обозначая сущность, похожую на элементы этого подмножества в (18), или типичный пример, эталон такого элемента (19–20).

Так, сказка — это не гипотеза, но упоминаемые в 18 гипотезы в своей неправдоподобности похожи на сказку. В таких случаях речь может идти о **сравнении**.

С другой стороны, липиды и аминокислоты — это примеры органических молекул, а Кальтенбруннер — эталон фанатика. В таких случаях мы будем говорить о **парадигматическом примере**.

В качестве ‘Y’ с лёгкостью может выступать не только одна или несколько именных групп (18–20), но и, к примеру, сочинённые группы инфинитива (21), цепочка клауз (22), сложное предложение (23).

(21) Старушка очень не любила / когда её беспокоили по *пустякам типа пострелять из пистолета или взорвать какой-нибудь объект*. [Сергей Дебижев. Два капитана II, к/ф, 1992]

(22) В одной из гробниц Древнего царства / то есть первая половина третьего тысячелетия / есть текст, где… ну там много.. / часто встречаются / мы знаем / в так называемых автобиографиях покойных / *фразы типа «Я давал хлеб голодному / я одевал нагого / я давал воду жаждущему»*… [Древнеегипетская книга мёртвых. Программа «Гордон» (НТВ), 2001]

(23) С другой стороны, жизнь относительно честных обществоведов тоже была несладкой: учащиеся, чуя слабину, доставали их самым неприличным образом, задавая *всякие неприятные вопросы типа: «А почему мы в социалистическом раю живём так хреново, а на прогнившем Западе джинсы и колбаса»*. [Константин Крылов. Мертворождённый // «Русская жизнь», 2012]

Как видно из (22–23), говорящие охотно используют в качестве типичного примера отрывки из чужой речи в том случае, когда специфицируемые сущности представляют собой некоторые отрезки дискурса — *фразы*, *вопросы*, *ответы* etc. Ни по структуре, ни по функции в тексте (22–23) принципиально не отличаются от (17–18). И в то же время связь между *типа* в таких конструкциях и частицей *типа* в роли ксенопоказателя представляется нам очевидной[18].

[18] Мы отграничиваем *типа* в составе параметрических конструкций от частицы *типа* на основании формальных признаков: пока в левом контексте присутствует специфицируемое существительное, мы вправе говорить о конструкции. На этом основании мы относим (22–23) к параметрическим конструкциям. Похожий случай представляет собой и английский пример 6:

(6) My mother would always say *stuff like: “Now, don’t go out without your coat!”*

В таких случаях наблюдается совмещение парадигматического примера и аппроксимативного цитирования.

С другой стороны, когда в состав конструкции входят в качестве специфицируемого элемента неопределённые местоимения (24–25)[19], она приобретает значение приблизительности.

(24) Это будет *что-то типа семи с чем-то детей на женщину.* [Далхат Эдиев. Арифметика населения. Проект Academia (ГТРК Культура), 2010]

(25) Потом возьмёшь такси до любого метро ближайшего / только не эту ветку. *Что-нибудь типа «Университет»* / ну / что угодно. [Константин Мурзенко. Апрель, к/ф, 2001]

Так, в (24) приводится приблизительная оценка количества, а в (25) подчёркивается нерелевантность приводимого примера — при соблюдении необходимых условий станция «Университет» вполне может быть заменена на другую (на *что угодно*).

Наконец, когда и ‘X’, и ‘Y’ представлены местоимениями, вся конструкция приобретает способность выражать субъективную оценку. Такова, например, широко распространённая конструкция *что-то типа того*, часто редуцирующаяся до *типа того* в разговорной речи:

(26) Сказала / что если бы вы были другом Старовойтовой / смысл был такой / если мне не изменяет память / очень много вопросов / не могу найти / но сейчас бы вы не были с ней рядом / *что-то типа того.* [Беседа А. Венедиктова с А. Илларионовым в эфире радиостанции «Эхо Москвы» // «Эхо Москвы», 2003–2004]

(27) [Лина] Ну вообще / это непроверенные данные.
[Андрей] М-м / понятно.
[Лина] Потому что… никто не знает точно.
[Андрей] Понятно / апокриф.
[Лина] Ну да / *что-то типа того.*
[Кира Муратова и др. Настройщик, к/ф, 2004]

(28) [Арсений] Я архитектор.
[Арина] Вы сочиняете дома́?
[Арсений] Ну *типа того.*
[Александр Кириенко и др. Инди, к/ф, 2007]

Дейктическое местоимение *того* отсылает к слову или фразе в левом контексте, оценивая их как неточные, но при этом вполне приемлемые на данном этапе развития дискурса. В частности, участник диалога получает возможность дать приемлемый с формальной точки зрения — пусть и уклончивый, почти пустой с точки зрения содержания — ответ, вместо того чтобы проигнорировать реплику собеседника.

[19] Или семантически опустошённые (*semantically bleached*) существительные, напр., *пустяки* в 21.

4. Дискурсивный маркер *типа*

4.1. ‘Пропозициональные’ употребления

В качестве дискурсивной частицы *типа* может вводить сравнение (29), характерный пример (30) или обозначение приблизительного количества, как в обсуждавшемся выше примере 16, который мы приводим здесь в сокращённом виде под номером (31): *типа* без труда можно заменить на *как* в (29), *например* в (30) и *примерно* в (31).

(29) Американец-то оказался круче всех крутых хохлов. Он у них там *типа новый Аль Капоне*. [Алексей Балабанов. Брат-2, к/ф, 2000]

(30) Между прочим / и Фройд в этом принимал отношение / и Куссмауль / там / ну / в общем / целый ряд весьма серьёзных имён можно назвать / которые стали один за другим открывать зоны мозга / ответственные за самые невероятные вещи / там. *Типа зоны / отвечающие за пение* / там */ за чтение вслух или чуть ли не за пользование пилочкой для крабов* / там. [Два мозга. Программа «Гордон» (НТВ), 2001]

(31) Слушай, ужас-то какой. Чё, прям совсем дорогой? *Типа* миллион? [Авдотья Смирнова, Анна Пармас. Кококо, к/ф, 2012]

4.2. Ксенопоказатели

Но и в том случае, когда *типа* используется в качестве ксенопоказателя, мы не можем однозначно утверждать, что говорящий оценивает цитату как неточную. Напротив, мы наблюдаем целый спектр возможных употреблений.

На одном полюсе этого спектра находятся случаи, которые вообще не подразумевают эпистемической оценки. Это либо обобщение частных высказываний (32), либо гипотетическая фраза, которая может быть произнесена при определённых обстоятельствах (33–34).

(32) Американцы / за что их / кстати / очень часто упрекают / *типа они там косные* / значит / *и не хотят знакомиться с другой культурой* / но восемьдесят процентов американцев отдыхают внутри Соединённых Штатов Америки. [Встреча президента РФ Д. Медведева со студентами факультета журналистики МГУ им. М. В. Ломоносова, 2012]

(33) Значит / для деревенского ролика нужен дед / коза / завалинка. Игорь Владимирович приобнимет деда / скажет там *типа* / *Не горюй* / *Иван Митрофаныч* / *поднимем село*. [Олег Фомин и др. День выборов, к/ф, 2007]

(34) Просто скажи / *ой* / *я* там *болела* / *типа у меня справка есть*. Чё-нить такое там. [Телефонный разговор студенток об учёбе // Из коллекции НКРЯ, 2015]

(32–34) близки к (22–23): и там, и там приводимое высказывание представляет собой частный случай типичного примера — примера фразы, характерной

или наиболее подходящей в описываемой ситуации. Невозможность эпистемической оценки в 32–34 поддерживается семантикой глагольной формы (хабитуалис / будущее время / императив).

В иных случаях, где речь идёт о конкретных событиях в прошлом и имеет место ситуация припоминания, оппозиция ‘достоверность / недостоверность’ как бы снимается: *типа* снимает с говорящего ответственность за точность формулировки[20] (35–36), но в то же время не исключает близкой к оригиналу и даже вполне точной передачи чужих слов (37).

(35) И в общем / короче / я распсиховалась / сказала то / что *типа отдава*… ну / *ничё не делайте с моим телефоном / я щас приеду и его заберу.* [Разговор подруг по телефону // Из коллекции НКРЯ, 2015]

(36) Там меня дядька посадил короче говоря / за машину // и спросил / *ну типа умеешь вообще кататься или нет // сидела / за рулём?* [Рассказ соседям о первом уроке вождения, 2007]

(37) Ну я искала / уже третью неделю ищу / короче / он мне пишет / *типа* / «*Вот / типа / в этом случае ты мне должна диск возместить*» / короче. [Разговор подруг // Из коллекции НКРЯ, 2009]

Напротив, в тех случаях, когда говорящий интерпретирует мысли или действия другого лица, комментируя их при помощи вымышленной фразы (*Interpretive Quotative*), аппроксимативная семантика маркера *типа* актуализируется.

(38) Ты просто думаешь / как… как тогда / что *типа вот нам не говорят / значит ничё делать не надо.* [Разговор подруг по телефону // Из коллекции НКРЯ, 2015]

(39) Ну / она [нрзб] когда отвечаю / то в окно смотрит там / то лицо такое делает / *типа я тупая*. Надоело… [Разговор студентки и школьницы об учёбе // Из коллекции Ульяновского университета, 2009]

4.3. Аппроксиматоры

Сказанное выше о ксенопоказателях применимо и к тем случаям, когда *типа* оформляет не цитирование, а просто описание ситуации. Аппроксимативная семантика ослабевает в случаях, подобных 40, где описывается типичная, обобщённая ситуация, и, наоборот, усиливается в ситуациях интеллектуальной работы и поиска (41), припоминания (42) или мыслительного затруднения, сомнения. Последний случай прекрасно иллюстрируется примером 1.

(40) Знаешь /как на выставках всяких ювелирные украшения демонстрируют? Их *типа на подушечку кладут* / чтоб лучше видно было и чтоб освещались равномерней… [Разговор подруг // Из коллекции НКРЯ, 2005]

[20] Ср. [Fleischman 1998] с. 38, [Подлесская, Стародубцева 2013] с. 32.

(41) Или / может быть / это бывает как некая вот идея / *типа* / *в смысле вот идея пространства / может быть*. [Беседа с преподавателем // Из коллекции НКРЯ, 2008]

(42) Я пришла в одну фирму / а там мужик был такой блатной такой.. ой-ой-ой / *там типа / там платили больше за час в два раза / чем обычно / тогда было сто рублей в час*. [Лида. Как я искала себе третью работу!!!, 2010]

Аппроксимативная семантика *типа* актуализируется в ситуации поиска подходящей номинации, при иронической номинации / описании и в случае нечёткости коммуникативных намерений.

4.3.1. Поиск подходящей номинации

В тех случаях, когда говорящие испытывают трудности с нахождением подходящей номинации, маркер *типа* позволяет сигнализировать о неточности выбранной номинации, используемой за неимением лучшего варианта. Так, в (43) говорящий пытается вспомнить слово *шулер*, а в (44) объясняет значение слова *смайлик* (‘значок, изображающий улыбку’): *типа улыбка* — это ‘то, что можно приблизительно обозначить словом улыбка’. В общем виде *типа N* означает ‘имеющий некоторые свойства N (но при этом отличающийся от него)’.

(43) У них там… ээ / был *типа* / ну / там… *листики дёргал / туда-сюда катаю / капусту вынимаю*… Ну там / вообще / на Ивана Дурко разводить мастер. [Константин Мурзенко. Апрель, к/ф, 2001]

(44) Смайлик / это значок. Ну *типа улыбка*.
[Алексей Попогребский. Как я провёл этим летом, к/ф, 2010]

4.3.2. Ирония

В (45) компонент значения ‘отличающийся от N’ трансформируется в ‘имеющий внешнее сходство с N, но противоположный N по своим сущностным свойствам’. *Типа честность* лишь внешне похожа на честность, а на самом деле представляет собой её полную противоположность. Такую трансформацию смысла можно с полным правом назвать иронией.

(45) Ты мне поэтому / значит / деньги за билеты отдавала / да? <…> Это типа честность такая. Благородство такое / да? Разводить / но не на деньги. [Алексей Учитель, Авдотья Смирнова. Прогулка, к/ф, 2003]

В (46–47) также можно усмотреть иронию, к которой здесь добавляется изобразительность, поддержанная вторым дискурсивным маркером *такой*.

(46) Ну вот / представляешь / подходит он ко мне такой *типа крутой. Пойдём / говорит / типа маленькая / потанцуем*. Ну и всякое такое. [Максим Пежемский, Константин Мурзенко. Мама, не горюй!, к/ф, 1997]

(47) Ну и вот / он такой / ой-ой / *у нас тут фирма / мы тут типа платим* / там / за… чтобы было качество хорошее. [Лида. Как я искала себе третью работу!!!, 2010]

4.3.3. Нечёткость коммуникативных намерений

Аппроксимативная семантика частицы *типа* в полной мере реализуется и в тех случаях, когда коммуникативные намерения говорящего сами по себе являются нечёткими либо расходятся с намерениями его собеседника.

Например, говорящий может сомневаться в уместности самого речевого акта. В этом случае *типа* смягчает категоричность высказывания, делая его более вежливым — неуверенная просьба в (48). В (49) при помощи маркера *типа*, говорящий уклоняется от точного ответа, скрывая от собеседницы истинное положение вещей.

(48) [Масяня] Это что / *я́* должна делать?
[Хрюндель] Ну / нет / ну… ну я думал / *типа* / *поможешь*. Хм! Ну… [Олег Куваев. Масяня, м/ф, 2002–2008]

(49) [Оля] [находит у Саши на одежде еловые иголки] Саш! За грибами что ли ходил?
[Саша Белый] Ну *типа да*.
[Алексей Сидоров, Игорь Порублев. Бригада, к/ф, 2002]

5. Фокус

Что касается фокусной функции, на материале НКРЯ и наблюдения над интернет-контентом мы можем сделать вывод, что фокусное употребление в чистом виде встречается весьма редко.

Так, в (50–51), так же как и в (40), *типа* предваряет новую актуальную информацию, следующую за темой ((51) можно легко преобразовать в (51а) с вынесением топика), в то время как аппроксимативная семантика несколько ослаблена: в (40) говорящая описывает, по-видимому, хорошо знакомую ей ситуацию, а в (50) говорящий вряд ли сомневается в точности своих слов. В то же время полностью исключать наличие аппроксимативной семантики в данных примерах мы не можем.

(50) Двойка у меня появилась после визита директрисы в СК. Я спросил тогда учителя по алгебре, *типа*, *имеете ли вы к этому отношение*?[21]

(51) А кашу же тоже / *типа* / варить надо.
[Разговор студенток о посте и о кризисе // Из коллекции НКРЯ]

(51а) А что касается каши, её тоже нужно варить.

Однако как в бытовом диалоге (52), так и в публичной устной речи (53–54) можно найти примеры, для которых фокусную функцию типа следует признать если не единственной, то по крайней мере, основной:

(52) Я не хотела на неё [на гусеницу] смотреть / и папа что-то такой: «Ну / иди / посмотри». А я так: «Ну / забей / забей…» Потом такой: «Ксения

[21] Пример (50) взят с сайта www.novayagazeta.ru.

/ иди сюда / иди сюда». Ну / я такая: «А-а-а-а…» и / *типа / чуть на неё не наступила*. [Разговор в компании подруг // Из коллекции НКРЯ, 2008]

(53) На встрече присутствовали / *типа / как молодые люди / так и ста… представители старшего поколения*. [Рассказ об общественной работе // Из коллекции НКРЯ, 2006]

(54) Ну сейчас / *типа / в нашем округе создан молодёжный совет*. [Рассказ об общественной работе // Из коллекции НКРЯ, 2006]

6. Заключение

По нашим наблюдениям, из трёх кластеров функций — ‘пропозициональных’, ‘текстовых’ и ‘экспрессивных’ — *типа* в составе параметрических конструкций тяготеет к выражению пропозициональных значений, в первую очередь сравнения и парадигматического примера, в то время как дискурсивный маркер *типа* свободно выражает все три кластера, тяготея при этом к экспрессивным употреблениям, причём доля ксенопоказателей среди последних очень велика. Доля же так называемых ‘текстовых’, или фокусных, употреблений, наоборот, крайне невелика, более того, они настолько редки в чистом виде, что мы рассматриваем их как отдельный тип значений, способный комбинироваться с экспрессивными, ср. **Таблицу 1**. В свете этих данных постулированный в работе [Fleischman, Yaguello 2004] переход от ‘пропозициональных’ употреблений к ‘экспрессивным’ через обязательную стадию ‘фокусных функций’ кажется маловероятным. Впрочем, последняя гипотеза требует дальнейшей проверки, возможно, на материале других корпусов, содержащих большую долю неформальной устной речи. Здесь же мы стремились ограничиться синхронным описанием.

Таблица 1. Соотношение различных употреблений частицы *типа* в устной речи на период 2008–2017 по данным НКРЯ

Экспрессивные	**Аппроксиматор**	**Ксенопоказатель**
	119	83
Фокусные	24	

Литература

1. *Арутюнова Н. Д.* (2000), Показатели чужой речи *де*, *дескать*, *мол*, Арутюнова Н. Д. (ред.) Язык о языке. М.: Языки русской культуры, с. 437–452.
2. *Левонтина И. Б.* (2010), Пересказывательность в русском языке, Доклады междунар. конф. «Диалог-2010». М.: Институт Проблем Информатики РАН, с. 284–288.
3. *Плунгян В. А.* (2008), О показателях чужой речи и недостоверности в русском языке: ‘мол’, ‘якобы’ и другие, B. Wiemer, V. A. Plungjan (eds.) Lexikalische Evidenzialitäts-Marker in slavischen Sprachen. Wiener Slawistischer Almanach, Sonderband 72. Munchen: Sagner, S. 285–311.
4. *Подлесская В. И., Егорова М. А.* (2017), От полнозначного имени к частице: аппроксиматор типа в зеркале корпуса с просодической разметкой, Е. В. Печенкова, М. В. Фаликман (ред.) Когнитивная наука в Москве: новые исследования. Материалы конференции 15 июня 2017 г. М.: Буки Веди, ИППиП, с. 283–288.
5. *Подлесская В. И., Стародубцева А. В.* (2013), О грамматике средств выражения нечёткой номинации в живой речи, Вопросы языкознания, № 3, с. 25–41.
6. *Русская грамматика* (1980), Шведова Н. Ю. (гл. ред.), М.: Наука.
7. *Dehé N., Stathi K.* (2016), Grammaticalization and prosody: the case of English *sort/kind/type of* constructions, Language 92 (4), Dec., pp. 911–947.
8. *Fleischman S.* (1998), Des jumeaux du discours, La Linguistique 34, № 2, pp. 31–48.
9. *Fleischman S.* (1999), Pragmatic markers in comparative and historical perspective: Theoretical implications of a case study. Paper presented at the 14th International Conference on Historical Linguistics, August 1999, Vancouver, Canada.
10. *Fleischman S., Yaguello M.* (2004), Discourse markers across languages, Carol Lynn Moder, Aida Martinovic-Zic (eds.). Discourse across languages and cultures. Amsterdam—Philadelphia: John Benjamins Publishing Company, pp. 129–147.
11. *Heine B., Kuteva T.* (2002), World lexicon of grammaticalization. Cambridge: Cambridge University Press.
12. *Savchenko D. S.* (2015), On Possible Functions of X_{ind} vrode/tipa/napodobie ‘like’ Y in Russian Colloquial Speech // Коммуникативные исследования. № 4(6), Омск: ОмГУ, с. 100–108.
13. *Underhill R.* (1988), *Like* is, like, focus, American Speech. Vol. 63, pp. 234–246.
14. *Yaguello M.* (1998), *Genre*, une particule d’un *genre* nouveau, Yaguello M. Petits faits de langue. Paris: Éditions du Seuil, pp. 18–24.

Источники

1. Национальный корпус русского языка www.ruscorpora.ru

References

1. *Arutjunova N. D.* (2000), Markers of indirect speech *de, deskat’, mol* [Pokazateli chuzhoj rechi *de, deskat’, mol*], Arutjunova N. D. (ed.) Language about language [Jazyk o jazyke], pp. 437–452.
2. *Dehé N., Stathi K.* (2016), Grammaticalization and prosody: the case of English *sort/kind/type of* constructions, Language 92 (4), Dec., pp. 911–947.
3. *Fleischman S.* (1998), Des jumeaux du discours, La Linguistique 34, № 2, pp. 31–48.
4. *Fleischman S., Yaguello M.* (2004), Discourse markers across languages Carol Lynn Moder, Aida Martinovic-Zic (eds.). Discourse across languages and cultures. Amsterdam—Philadelphia: John Benjamins Publishing Company, pp. 129–147.
5. *Fleischman S.* (1999), Pragmatic markers in comparative and historical perspective: Theoretical implications of a case study. Paper presented at the 14th International Conference on Historical Linguistics, August 1999, Vancouver, Canada.
6. *Heine B., Kuteva T.* (2002), World lexicon of grammaticalization. Cambridge: Cambridge University Press.
7. *Levontina I. B.* (2010), Quotation and rendering markers in Russian [Pereskazyvatelnost’ v russkom jazyke], Papers from Annual International Conference “Dialogue”, pp. 284–288.
8. *Plungian V. A.* (2008), On markers of indirect speech and unreliability [O pokazateljax chuzhoj rechi i nedostovernosti v russkom jazyke: *mol*, *jakoby* i drugie], Wiemer B. & V. A. Plungjan (Hrsg.), Lexikalische EvidenzialitätsMarker in slavischen Sprachen [Wiener Slawistischer Almanach, Sonderband 72], Sagner, München, S. 285–311.
9. *Podlesskaya V., Egorova M.* (2017), From a full-fledged noun to a particle: the Russian approximator *tip* viewed through the lense of prosodically annotated corpora, Cognitive Science in Moscow, pp. 283–288.
10. *Podlesskaya V. I., Starodubtseva A. V.* (2013), On grammar of hedges in Spoken Language // Issues in linguistics [Voprosy jazykoznanija], № 3, pp. 25–41.
11. *RG-80* — Russkaya grammatika [Russian grammar]. Shvedova N. Yu. (ed.). Moscow, Nauka, 1980.
12. *Savchenko D. S.* (2015), On Possible Functions of X_{ind} vrode/tipa/napodobie ‘like’ Y in Russian Colloquial Speech // Communicative Studies [Kommunikativnye issledovaniya], № 4(6) Omsk: OmSU, pp. 100–108.
13. *Underhill R.* (1988), *Like* is, like, focus, American Speech. Vol. 63, pp. 234–246.
14. *Yaguello M.* (1998), *Genre*, une particule d’un *genre* nouveu, Yaguello M. Petits faits de langue. Paris: Éditions du Seuil, pp. 18–24.

Sources

1. Russian National Corpus www.ruscorpora.ru

A STUDY OF MACHINE LEARNING ALGORITHMS APPLIED TO GIS QUERIES SPELLING CORRECTION

Fomin V. V. (wadimiusz@gmail.com)
Novosibirsk State University, Novosibirsk, Russia

Bondarenko I. Yu. (i.yu.bondarenko@gmail.com)
2GIS, Novosibirsk, Russia

The problem of spelling correction is crucial for search engines as misspellings have a negative effect on their performance. It gets even harder when search queries are related to a specific area not quite covered by standard spell checkers, such as geographic information systems (GIS). Moreover, standard spell-checkers are interactive, i.e. they can notice a misspelled word and suggest candidate corrections, but picking one of them is up to the user. This is why we decided to develop a spelling correction unit for 2GIS, a cartographic search company. To do this, we have extracted and manually annotated a corpus of GIS lookup queries, trained a language model, performed various experiments to find the best feature extractor, then fitted a logistic regression using an approach suggested in SpellRuEval, and then used it iteratively to get a better result. We have then measured the resulting performance by means of cross-validation, compared at against a baseline and observed a substantial increase. We also present an interpretation of the result achieved by calculating and discussing the importance of specific features and analyzing the output of the model.

Keywords: spell checker, geographic information system, language model, text corpus, local search

ИССЛЕДОВАНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ИСПРАВЛЕНИЯ ОПЕЧАТОК В ЗАПРОСАХ К ГИС

Фомин В. В. (wadimiusz@gmail.com)
Новосибирский государственный университет, Новосибирск, Россия

Бондаренко И. Ю. (i.yu.bondarenko@gmail.com)
2ГИС, Новосибирск, Россия

1. Introduction

The problem of spelling correction has a long history of research. Works in this area generally tend to belong to either of the following fields:

1. candidate search, i. e. suggesting corrections for a typo, such as the pioneer work of the field [Damerau 1964];
2. candidate selection, i. e. evaluating the candidate suggestion quality to select the best suggestion, such as [Kernighan, Church, Gale 1964].

Lately, there has been a research into the problem of context-sensitive spell-checking, i. e. using the word context for candidate selection, for instance [Golding, Roth 1999]. Some try to embed various approaches into a single candidate selection system, such as [Gao et al. 2010].

However, both tasks are rather challenging if the spell check algorithm is designed for specific use, e. g. for geographic information system (GIS) lookups such as electronic maps developed by Russian cartographic company 2GIS. Candidate search, on one hand, may pose a problem as standard spell checkers often appear unaware of GIS-related words, such as names of streets, companies etc, such as in the following example, where Hunspell treated a correct query as a typo and suggested this correction, undoubtedly because this tool is too general:

(1) *хоум кредит банк → ухом кредит банк*
(The part before "→" represents the original query,
the part after "→" represents the suggestion).

On the other hand, choosing a candidate may also be a problem because of the specific language used in the area. Search queries differ drastically from other forms of language, as syntactic and morphological information are much less helpful than usual.

Although there are papers describing spelling correction of search queries, such as [Martins, Silva 2004] for candidate search and [Wilbur, Kim, Xie 2004] for candidate ranking, technical details of most search engine spell-checkers are a trade secret, which makes the problem under discussion even more challenging.

We decided to develop a unit aimed specifically at correcting search query typos. This means that we had to create a corpus for supervised learning, design a feature extractor (which included training a language model), design and fit a spell-checking model, and evaluate its performance.

2. Related work

One of the oldest works related to correcting spelling errors is [Blair 1960]. It introduces the idea of spelling corrections based upon a list of correctly spelled words and a string metric. Although the string metric suggested in this work gained no popularity, the approach itself is crutial to the task of spelling correction.

The two classical works for the task are [Damerau 1960] and [Levenshtein 1966], upon which the most popular string metric, the so-called Damerau–Levenshtein

distance, is based. The distance between two words is the number of operations it takes to turn one word into the other one.

The problem of candidate search is considered in [Brill, Moore 2000], a paper which discusses the noisy-channel model, a way of determining which correction is better for a certain typo by assigning each typo a probability and using it to determine the score of each correction. This, however, means that the model is incapable of taking the context into account and deciding whether or not a correction suits its context.

An interesting approach to spelling correction of search queries is presented in [Cucerzan, Brill 2004]. The authors of this paper suggest a spell-checker designed for search queries should rely upon the statistcs of search queries and correct them iteratively, in several steps, making a typo less malign wth each iteration. This is opposed to the usual approach that relies upon a list of correct words and suggests correcting a typo in one move.

3. Text corpus

We have used two query corpora to fit our model. We refer to the first corpus as "supervised" and to the second as "unsupervised". The first corpus consists of 14,400 manually annotated queries. Each entry in this corpus looks like this:

(2) *большой сухаревский переклок 23/25,большой сухаревский переклок 23 25,большой сухаревский переулок 23 25,1*

The entries contain four fields. The first field is the original query, the second field represents the result of a basic preprocessing, the third one is a gold-standard correction, and the fourth one is a binary classification label, where "1" stands for "This query contains a typo" and "0" stands for "This query does not need being corrected".

This corpus is used for supervised learning and for evaluating its performance metric.

The second corpus contains around one million raw queries. This corpus was used for creating a language model.

4. Model

Our model can be accessed at [Fomin 2017].

Our suggested correction lookup unit was influenced by the work of Cucerzan and Brill [Cucerzan, Brill 2004]. The approach presented in this work (and implemented in our model) is as follows. When we use dictionaries that only consist of properly spelled words for candidate search task, we have to pick from two extremities: only considering corrections that are close to the original word or including corrections that are rather distant from the analyzed word.

In the first case we are bound to fail in any complicated case like

(3) *anol scwartegger*

instead of

(4) *arnold schwarzenegger*

In the second case we have to process a stupendous number of corrections, some of which may be erroneously taken for the right ones, which also leads to poor model performance. The solution is to take misspelled words into account when performing the candidate search, and to make the spelling correction model process the sentence several times iteratively, so that every time the misspelling becomes less malign:

(5) *anol scwartegger → arnold schwartnegger → arnold schwarznegger → arnold schwarzenegger*

Our candidate ranking system was inspired by the work of Sorokin and Shavrina. [Sorokin, Shavrina 2016] We have created a feature extractor which takes in two strings (one of them being the original query, the other the suggested correction) and returns the following features:

1. Length of the correction.
2. Simple Levenshtein distance between the original query and the correction.
3. The score of the correction evaluated by a SRILM [Stolcke 2002] ngram-model.
4. The number of out-of-vocabulary words in the corrections.
5. The number of vocabulary words that became out-of-vocabulary after the correction.
6. The number of out-of-vocabulary words that became vocabulary words after the correction.
7. The number of words whose correction is more frequent than the original correction.
8. Simple Levenshtein distance between the original tokens and their corrections, where the original tokens are only considered if they are out-of-vocabulary.
9. Simple Levenshtein distance between the original tokens and their corrections, where the original tokens are only considered if they are in the vocabulary.
10. The number of original tokens whose corrections are 1 Levenshtein operation far from the originals.
11. The number of space symbol deletions.
12. The number of space symbol insertions.
13. The number of out-of-vocabulary words that can be split into two vocabulary words.
14. Weighted Levenshtein distance where the weight of a substitution operation is the distance of the corresponding symbols on a phone keyboard layout.
15. Weighted Levenshtein distance where the weight of a substitution operation is the distance of the corresponding symbols on a phone keyboard layout and the weight of a deletion operation is 10.
16. Weighted Levenshtein distance where the weight of a substitution operation is the distance of the corresponding symbols on a phone keyboard layout and the weight of a insertion operation is 10.

17. Simple Levenshtein distance between the phonetic codes (described in [Sorokin, Shavrina 2016]) of the original query and the suggestion.
18. The number of corrections of the form "*а → о*, *о → а*, *е → и*, or *и → е*".
19. The number of corrections of the form "*ы → и*, *ё → о*, *ю → у* after *ж*, *ч*, *ш*, *щ*".
20. The number of corrections of the form "*цы → ци*" or "*ци → цы*".
21. The number of corrections of the form "*ыва → ова*".
22. The number of corrections of the form "*аро → оро*" or "*ало → оло*".
23. The number of corrections of the form "*э → е*".
24. The number of corrections of the form "*ща → ще*".
25. The number of corrections of the form "*пре → при*" or "*при → пре*".
26. The number of corrections of the form "*э → и*".
27. The number of corrections of the form "*ё → йо*" or "*е → йо*".
28. The number of corrections of the form "unvoiced consonant → its voiced equivalent", or "unvoiced consonant → its voiced equivalent".
29. The number of corrections of the form *"зн → здн"*, *"сн → стн"*, "*сл → стл*", "*нст → нтст*", *"здн → зн"*, *"стн → сн"*, *"стл → сл"*, or "*нтст → нст*".
30. The number of corrections of the form *"хк → гк"*.
31. The number of corrections of the form "*н → нн*", "*с → сс*", "*м → мм*", "*ф → фф*", or vice versa.
32. The number of corrections of the form "ь → ъ" or "ъ → ь".
33. The number of corrections of the form "insertion of ь as the fourth-to-last letter".
34. The number of corrections of the form "*тся → ться*" or "*ться → тся*".

We have used SRILM [Stolcke 2002] and our untagged corpus to fit a language model used in the feature extractor as the feature #3. After we tried running our spell checker with models at different smoothing types and context windows and found that the best performance is achieved when a bigram model with Kneser-Ney smoothing is used. The fact that no bigger order of ngram-model (trigrams etc.) was required is remarkable and somewhat characteristic of the language of search queries.

Features from 18 to 34 were included in order to capture some typical cases of typos, so that the suggestions that assume a more usual typo would be considered a better corrections by the model.

We have also tried using several morphological or simple syntactic models. Attempting to train our model to generalize morphological properties of the corpus, we trained a SRILM model on a version of the "unsupervised" corpus with words replaced by morphological tags, making TreeTagger [Schmid 2013] do the tagging; we also tried to replace all the words but the most frequent ones. We tried to make our model understand simple syntax (such as the fact that a word in dative case is expected after the preposition к) by leaving prepositions intact and replacing other words with morphological tags. We gave up these ideas as they did not improve the model performance.

We have also found that it is better to replace numbers in addresses of building by an artificial token, like this:

(6) *семёновская 9 → семёновская <NUMBER>*

This is because these numbers are not informative in spell-checking and have a bad effect on the model performance.

After features are extracted, we create the training set using the approach described in [Sorokin, Shavrina 2016]. For each query, we have one "winner suggestion" and a lot of "loser suggestions". We then compute the difference between the "winner suggestion" and each of the "loser suggestions" and assign them the label "1". Then, we take the same vectors multiplied by −1 and assign them with the label "0". We then transmit the resulting training set to a logistic regression model. By doing so, we train this model to maximize feature weights if the feature indicates that the suggestion is good and minimize feature weights of features that indicate that the suggestion is bad. This approach is equivalent to the one used in [Sorokin, Shavrina 2016], and it appeared very helpful.

After the training is done, we take the test set, generate a list of suggestions, put it through the trained logistic regression model, and claim that the suggestion with the highest score is the winner. After this is done, the winner sentence itself is now treated as the misspelled sentence which is to be corrected. This is repeated iteratively until the model recognizes that no further improvement can be done:

(7) *краснопресн* → *краснопресне* → *краснопресненеская* → *краснопресненская*

5. Evaluation

Evaluation in machine learning tasks consists of two main steps, namely picking and computing an evaluation metrics and comparing the results of the system in question against a baseline algorithm. When evaluating a spell-checking algorithm, both steps pose a certain difficulty.

First, as discussed in [Sorokin et al. 2016], typical spelling correction performance metrics such as the percentage of correctly processed sentences (or, in our case, queries) are not representative enough because they do not take the number of misspellings and corrections into account, and also because they make no difference between various cases such as making an undesired correction or leaving out a misspelled word that was to be corrected, so a model which performs no corrections at all would still have a 90% sentence accuracy upon a corpus with 10% misspelled sentences.

Second, comparing the model against a baseline would require the output of an algorithm on our corpus which is difficult because autonomous spell-checkers designed for search engine queries are not open-source systems, or available for researchers; typically, nor are their technical details.

We have treated these problems as follows.

For evaluating our model, we have implemented the same approach as the organizers of SpellRuEval described in [Sorokin et al. 2016]. Instead of using primitive performance metrics, we treated each word or word group that was and/or should have been corrected as a separate case, marking them as true positives, false negatives or false positives. Cases where the original query had contained a typo and was corrected by the system in the desired way were considered to be true positives. Instances where the typo was not affected by the spell checking system, or was corrected in a way

different from the golden standard were treated as false negatives. Occasions when a properly spelled word was mistaken for a misspelling were treated as false positives.

We have then evaluated precision, recall and F1-measure using standard formulas for binary classification. We have performed a cross-validation over 5 folds to do so.

We have used Hunspell as a baseline, with the following results:

	F1-score	Precision	Recall
Our algorithm	73.5%	94.8%	58.6%
Hunspell	20.8%	19.6%	22.2%

Details of the evaluation are presented below.

Hunspell. First, we have put our corpus through Hunspell [Németh 2003], a widespread interactive spell-checker, by asking it word by word whether the given word is correct and replacing the supposed misspellings by the best of all suggestions Hunspell could make (if any). It is important to note that it was not our goal to test our dictionary against Hunspell's standard dictionary; we wanted to find out whether we succeeded in taking the language of the corpus into account and if our language modeling actually resulted in an improvement. Thus, we updated its standard list of correct tokens with a list of all streets, names of enterprises, cafés etc, so that we are sure that the difference in models' performance, if we observe any, is a caused by the difference in approaches to candidate selection and not in the dictionaries. A comparable result in Hunspell and our model would mean that language modeling was an unsuitable approach for correcting GIS search queries; observing our model's performance exceed qualities would mean that modeling the language of search queries for spelling correction is a good idea. Since Hunspell sorts the candidate corrections from most to least probable, we considered the first suggestion to be Hunspell's choice. The gap between the results of the two systems presented above is an evidence that language modeling is crucial for correcting the spelling of search queries and that our attempt to implement a system able to perform language modeling for spelling correction of search queries was successful.

SpellRuEval. We could not properly compare our results with the results of SpellRuEval winners [Sorokin, Shavrina 2016] since the source code is not available for us; however, it is interesting to note that they demonstrate an F1-score of 75%, which is a comparable result. It is hard to draw any conclusions from that because this score was achieved on quite a different corpus. On one hand, the lack of grammatical information and the amount of unusual tokens make the spell-checking of search queries a harder task. On the other hand, the plenty of morphological and syntactic information in regular texts enlarges the number of candidates and makes the candidate selection a more subtle task. In general, the fact that the winners of SpellRuEval achieved the result of 75% suggests that it is an acceptable result, although this is much less of a solid evidence than the baseline algorithms discussed above.

6. Analysis

After the training, we have extracted the feature weights from our logistic regression to know how much each feature contributes to the final decision of the model. A huge absolute value of a feature weight means that it plays an important role in taking decisions (a large positive coefficient means that examples with larger values of this feature tend to belong to class "1", whereas a large negative weight means that examples with larger value of this feature belong to class "0"). The result is presented on the following bar chart:

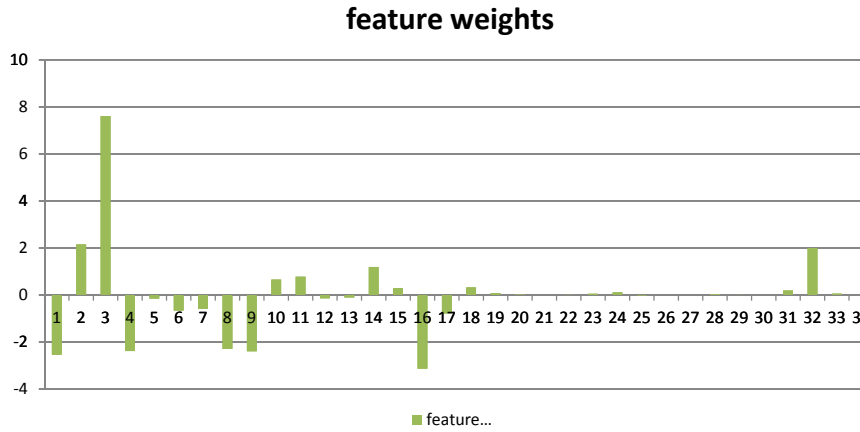


Fig. 1. The weights of various features in our feature extractor; each bar has a number and represents the feature of the same number; the height of a bar represents the weight of the corresponding feature

Here, each number represents a certain feature; for more clarity, we briefly repeat the descriptions of features whose detailed versions are available in part 3 "Model" (OOV stands for out-of-vocabulary):

1	Correction length	18	*а* → *о*, *о* → *а*, *е* → *и*, or *и* → *е*
2	Simple Levenshtein distance	19	*ы* → *и*, *ё* → *о*, *ю* → *у* after *ж*, *ч*, *ш*, *щ*
3	Ngram-model score	20	*цы* → *ци* or *ци* → *цы*
4	OOV in corrections	21	*ыва* → *ова*
5	Vocabulary → OOV	22	*аро* → *оро* or *ало* → *оло*
6	OOV → vocabulary	23	*э* → *е*
7	Corrections that are more frequent than the originals	24	*ща* → *ще*
8	Simple Levenshtein distance, only OOV originals	25	*пре* → *при* or *при* → *пре*
9	Simple Levenshtein distance, originals in vocabulary only	26	*э* → *и*

10	1-operation corrections	27	*ё* → *йо* or *е* → *йо*
11	Space deletions	28	unvoiced → voiced, or voiced → unvoiced
12	Space insertions	29	*зн* → *здн*, *сн* → *стн*, *сл* → *стл*, *нст* → *нтст*, *здн* → *зн*, *стн* → *сн*, *стл* → *сл*, or *нтст* → *нст*
13	OOV that can be split into two vocabulary words	30	*хк* → *гк*
14	Weighted keyboard layout Levenshtein distance	31	*н* → *нн*, *с* → *сс*, *м* → *мм*, *ф* → *фф*, or vice versa
15	Weighted Levenshtein distance with insertion weight 10	32	*ь* → *ъ* or *ъ* → *ь*
16	Weighted Levenshtein distance with deletion weight 10	33	insertion of *ь* as the fourth-to-last letter
17	Simple Levenshtein distance between phonetic codes	34	*тся* → *ться* or *ться* → *тся*

It is worth noting that the ngram-model score, represented by the feature #3 on Fig. 1 above, is by far the most important component of the feature vector. A high estimated probability of a query correction based upon the language of the corpus as a is an important evidence that the correction under discussion is to be accepted.

Other features with a high positive weight include feature #32, which will be discussed later, and, strangely, feature #2 which represents the simple Levenshtein distance between the original and the correction. This may be because there are several string metric features in the corpus, sometimes with close results, so that their contribution is shared and the weight of this specific feature is somewhat dependent on the initialization.

Other features of importance have a negative weight which means that they are useful for rejecting bad suggestions, rather than selecting a good one. This includes feature #1, which represents correction length and suggests that shorter corrections are better; feature #4 which means the number of out-of-vocabulary words in the correction; and features #8, #9, and #16, representing various string metrics. It is also interesting to note that feature #16 that signifies weighted Levenshtein distance with substitutions weighted proportionally to the distances on a keyboard layout and deletions weighted 10 times as high as insertions, turned out to be the most successful string metric. This might be because it represents the properties of search queries where substitutions are caused by the so called fat finger syndrome (mistakingly hitting adjacent keys on a keyboard) and typos that delete a character are much more common than those that insert one.

It is also interesting that features from #18 to #34 that were designed to represent typical typo cases mostly did not work out. The only feature of this type that has a clearly high absolute value of the weight is #32, which treats cases like:

(8) *сьесть* → *съесть*

This is due to the fact that letter ъ, on most phone keyboard layouts, can only be accessed via a certain manipulation with the key ь. Other features have a relatively low importance, either because they are represented by other features already, such as feature #18 vs. feature #17 for instances like

(9) *улеца → улица,*

or because it represents a relatively rare type of misspellings not quite relevant for the corpus, such as feature #29 for cases like

(10) *агенство → агентство.*

Besides, we have analyzed the output of our model to find the kinds of incorrect behavior that the described system is prone to. Although no particular regularities are evident in the output, it could be inferred from the discussed data that the most complicated cases that lead our system to false negative result are those whose originals are rarely seen in the search queries, such as

(11) *осташелвскпя → осташелвскпя*

instead of

(12) *осташелвскпя → осташевская,*

and queries chopped off at the middle of the string:

(13) *залес → залесс*

instead of

(14) *залес → залесского.*

7. Acknowledgements

The authors are grateful to Alexander Krinitsyn for his valuable advice. We also thank 2GIS who provided us with a corpus of GIS search queries and Botan Investments for supporting students' interest to machine learning.

References

1. *Blair C. R.* (1960), A program for correcting spelling errors. Information and Control. Vol. 3, №1, pp. 60–67.
2. *Brill E., Moore R. C.* (2000). An improved error model for noisy channel spelling correction. In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (pp. 286–293). Association for Computational Linguistics.
3. *Cucerzan S., Brill E.* (2004), Spelling correction as an iterative process that exploits the collective knowledge of web users, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Copenhagen.

4. *Cucerzan S., Brill E.* (2004). Spelling correction as an iterative process that exploits the collective knowledge of web users. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing.
5. *Damerau F. J.* (1964), A technique for computer detection and correction of spelling errors, Communications of the ACM, Vol. 7, № 3, pp. 171–176.
6. *Fomin V.* (2017), A term project on spell-checking. [ONLINE] Available at: https://github.com/wadimiusz/spellchecker [Accessed 20 February 2018].
7. *Gao J. et al.* (2010), A large scale ranker-based system for search query spelling correction, Proceedings of the 23rd International Conference on Computational Linguistics, Beijin, pp. 358–366.
8. *Golding A. R., Roth D.* (1999), A winnow-based approach to context-sensitive spelling correction, Machine learning, Vol. 34, № 1–3, pp. 107–130.
9. *Kernighan M. D., Church K. W., and Gale W. A.* (1990), A spelling correction program based on a noisy channel model, In Proceedings of the 13th conference on Computational linguistics, Avignon, pp. 205–210.
10. *Martins B., Silva M. J.* (2004), Spelling correction for search engine queries, Advances in Natural Language Processing, Springer, Berlin, Heidelberg, pp. 372–383.
11. *Németh L.* (2003) Hunspell. [ONLINE] Available at: http://hunspell.github.io/ [Accessed 20 February 2018].
12. *Schmid H.*(2013) Probabilistic part-of-speech tagging using decision trees, New methods in language processing, Manchester, p. 154.
13. *Sorokin A. A. et al.* (2016), Spellrueval: the first competition on automatic spelling correction for Russian, Proceedings of the Annual International Conference "Dialogue", Moscow.
14. *Sorokin A. A., Shavrina T. O.* (2016), Automatic spelling correction for Russian social media texts, Proceedings of the International Conference "Dialog", Moscow, pp. 688–701.
15. *Stolcke A.* (2002), SRILM-an extensible language modeling toolkit, Seventh international conference on spoken language processing, Denver.
16. *Wilbur W. J., Kim W., Xie N.* (2006), Spelling correction in the PubMed search engine, Information retrieval, Vol. 9, № 5, pp. 543–564.

DISCOVERING AND ASSESSING HEATED ARGUMENTS AT THE DISCOURSE LEVEL

Galitsky B. (boris.galitsky@oracle.com)[1,2],
Taylor R. (ray.taylor@oracle.com)[1]
[1]Oracle Inc., Redwood Shores CA, USA
[2]Higher School of Economics, Moscow, Russia

The problem of detecting heated arguments in text such as political debates and customer complaints is formulated as tree kernel learning of discourse structures. Affective argumentation structure is discovered in the form of discourse trees extended with edge labels for communicative actions. Extracted argumentation structures are then encoded as defeasible logic programs and are subject to dialectical analysis, to establish the validity of the main claim being communicated. We evaluate the accuracy of each step of this affect processing pipeline as well as overall performance.

1. Introduction

When an author attempts to provide a logical or affective argument in for something, a number of argumentation patterns can be employed. One of such patterns is to make a claim emotionally loaded, heated, escalated, associated with a confrontation. In text, argumentation patterns are associated with certain discourse features, and heated arguments are expressed in text by discourse means attempting to amplify the strength of these arguments.

The basic points of argumentation are reflected in rhetoric structure of text. A text without an argument, with a heated argument and with a logical one would have different rhetoric structures (Moens et al., 2007). When an author uses an affective argument instead of logical, it does not necessarily mean that his argument is invalid [Galitsky et al 2009). The goal of this study is to explore when a heated argumentation is valid. We introduce the notion of *heated argumentation* to circumscribe a special class of argumentation associated with strong emotions and sentiments.

Frequently people say of a politician's speech, *Oh, that's just rhetoric*, assuming that the words of politicians are empty verbiage or hot air. Frequently politicians do their most to sound impressive but indeed are saying nothing with real meaning. Sometimes politicians are making promises his listeners believe he has no intention of keeping. The use of rhetoric in an intuitive sense in speeches: both bad, dishonest and good ones is only the most visible use of rhetoric. In this work we attempt to treat the intuitive notion of rhetoric computationally with a special focus on heated rhetoric. We expect the strongest, heated arguments to have a more prominent underlying rhetoric structure.

We select the Rhetoric Structure Theory (RST, [Mann and Thompson 1988]) as a means to represent discourse features associated with heated argumentation.

Nowadays, a performance of both rhetoric parsers and argumentation reasoners has dramatically increased, and a discourse structure of text to be learned is formed from text automatically (Galitsky 2017a). Taking into account the discourse structure of conflicting dialogs, one can judge on the authenticity and validity of these dialogs in terms of validity of heated argumentation. In this work we will evaluate the *combined* argument validity assessment system that includes both the *discourse structure extraction* and *reasoning about it* with the purpose of validation of the complainant's claim.

Most of the modern techniques treat computational argumentation as specific discourse structures and perform detection of arguments of various sorts in text, such as classifying text paragraph as argumentative or non-argumentative ([Sardianos et al., 2015], [Stab and Gurevych, 2014], [Bondarenko, et al., 1997]). In this paper we intend to build the *whole heated argumentation pipeline*, augmenting argument extraction from text with its logical analysis. This pipeline is necessary to deploy an argumentation analysis in a practical decision support system:

1) Extract syntactic features;
2) Compute segmentation into elementary discourse units;
3) Build discourse trees;
4) Label their nodes with extracted communicative actions;
5) Form logical representation for clauses extracted from discourse tree;
6) Identify the main claim;
7) Given the logical representation as a Defeasible Logic Program, confirm or reject the main claim;
8) Produce a decision on whether argumentation for the main claim is acceptable or not.

Building this pipeline, we leverage two research areas: argument-mining, which is a linguistic-based, and logical validation of an argument, which is logic based. To the best of our knowledge, nowadays the former research area supports extracting various kinds of arguments from text on a scale, and the latter research area focuses on logical argumentation analysis of limited manually constructed argumentation structures. The contribution of this paper, the pipeline which implements the algorithms discovered in both these research areas, allows to perform logical analysis of a high quantity of heated arguments extracted from text. Therefore, industrial applications of mining and reasoning about heated arguments become possible. Since the paper combines linguistic and logical analyses, knowledge of both these domains is required from the reader to follow the whole pipeline of understanding heated arguments.

The concept of automatically identifying argumentation schemes was first discussed in (Walton et al., 2008). In (Ghosh et al., 2014) authors investigate argumentation discourse structure of the specific type of communication—online interaction threads. Identifying argumentation in text is connected to the problem of identifying truth, misinformation and disinformation on the web (Pendyala and Figueira, 2015, Galitsky 2015, Pisarevskaya et al 2015). In (Lawrence and Reed, 2015) three types of argument structure identification are combined: linguistic features, topic changes and machine learning.

To represent the linguistic features of text, we use the following sources:

1) *Rhetoric relations* between the parts of the sentences, obtained as a *discourse tree*.
2) *Speech acts, communicative actions*, obtained as verbs from the VerbNet resource (the verb signatures with instantiated semantic roles).

To assess the logical validity of extracted argument, we apply Defeasible Logic Program (DeLP, Garcia and Simari 2004), part of which is built on the fly from facts and clauses extracted from these sources. We integrate heated argumentation detection and validation components into a decision support system that can be deployed, for example, in a Customer Relationship Management (CRM) domain. To evaluate our approach to extraction and reasoning about argumentation, we choose the dispute resolution / customer complaint validation task because complainants frequently use heated argumengts to bring their point across. Most complainants are in a strong emotional distress due to a disparity between what they expected and what they received. Moreover, heated arguments appear in response to how companies communicate the issues with complainants. Most complaint authors report incompetence, flawed policies, ignorance, indifference to customer needs and misrepresentation from the customer service personnel. The complainants have frequently exhausted conventional communicative means available to them, confused, seeking recommendation from other users and advise others on avoiding particular financial service. Multiple affective argumentation patterns are used in complaints; the most frequent is an intense description by a complainant on a deviation of what has actually happened from what was expected, according to a common sense. This pattern covers both valid and invalid argumentation.

2. Representing Discourse for Heated Argumentation

We provide an example of conflicting agents providing their interpretation of certain events.

We show an example of a discourse tree (DT) for a heated argumentation of a customer treated badly by a credit card company American Express (amex) in 2007 (Fig. 1). Text split into logical chunks is as follows:

[I 'm another one of the many][that has been carelessly mistreated by American Express .] [I have had my card since 2004 and never late .] [In 2008][they reduced my credit limit from $16,600 to $6,000][citing several false excuses .] [Only one of their excuses was true—other credit card balances .] [They also increased my interest rate by 3 %][at the same time .] [I have never been so insulted by a credit card company .] [I used to have a credit score of 830 , not anymore , thanks to their unfair credit practices .] [They screwed my credit score .] [In these bad economic times you 'd think][they would appreciate consistent paying customers like us][but I guess][they are just so full of themselves .] [I just read today][that their CEO stated][that they will be hurt less than their competitors][because 80 percent of their revenues][are generated from fees.That][explains their callous , arrogant , unacceptable credit practices .] [It seems][they have to screw every cardholder][they can before the new law becomes effective .] [Well America , let 's learn from our appalling experience][and stop using our American Express credit card][so we can pay it off !].

We first explain how a traditional discourse tree encodes information flow in a paragraph of text. Text is split into logical chunks (elementary discourse units, EDUs) according to the order the entities of text are being introduced, attributes attached to them, and inter-relationships established. The author first introduces her opponent, describes how this opponent treats herself and others unfairly (according to her viewpoint) and enumerates different steps of this treatment.

DT is a tree where the EDUs are the labels of the terminal nodes. The other nodes are labeled with rhetorical relations encoding the type of logical links between the EDUs, such as *elaboration* (default), *attribution, cause* and others. Rhetorical relations hold not only between EDUs but also between the higher level logical chunks which might in turn include the lower level ones. That is how logical flow of text such as a heated argument can be visualized hierarchically. At the highest level, this text is split into two parts following the presentation sequence: 1) what happened; 2) what I think about it. This presentation style is covered by the rhetorical relation *topic-comment* (shown at the top of hierarchy).

In this study, to demonstrate the discourse features associated with heated argumentation, we augment the information on logical flow which is encoded in a traditional discourse tree and extend it by two components:

1) Communicative actions, showing how some elementary discourse units are being communicated [Galitsky 2017b];
2) Sentiment associated with some elementary discourse units.

It turns out that to differentiate a heated argumentation from a default, logical argumentation, 1) and 2) are essential. We refer to an extension of DT as a Communicative DT, CDT. CDT is a DT with labels for edges that are the VerbNet expressions for verbs (which are communicative actions, CA). Arguments of verbs are substituted from text according to VerbNet frames. The first and possibly second argument is instantiated by agents and the consecutive arguments—by noun or verb phrases which are the subjects of CA.

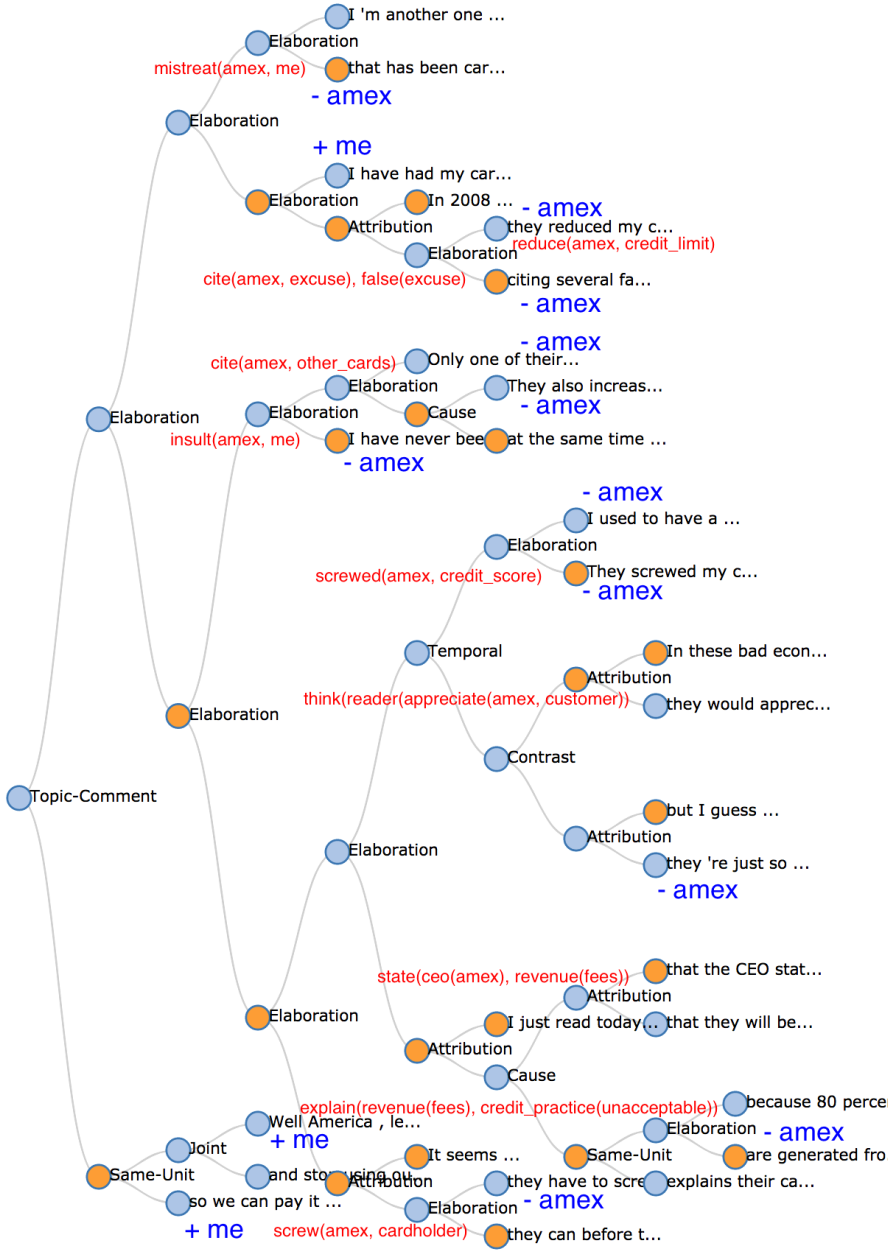


Fig. 1: A communicative discourse tree that includes labels for communicative actions and sentiments. Visualization of [Joty et al 2013] is used

In Fig. 1, the verbs communicative actions such as *mistreat(amex, me)* augment DT with necessary information about the text to match with other similar DTs. The

sequence of communicative actions provides information on the structure of a dialogue between a proponent and an opponent of a given argument. This information is complementary to what DT encodes for logical chunks provided irrespectively of how the entities from these chunks were communicated. Communicative actions are labels of the edges of the DT leading to the terminal nodes; sentiments are labels of these edges as well. We denote a sentiment polarity as + or—and the subject of this sentiment as the proponent (*me*) and opponent (here, *amex*). Naturally, an author provides an argument for how she is right and hew opponent is wrong, therefore one expects the positive sentiments with the author's EDUs and the negative ones with her opponents' EDUs. For each label, it is attached to the CDT edge nearest to the right end of the label expression.

Argumentation analysis needs a systematic approach to learn associated discourse structures. The features of CDTs could be represented in a numerical space argumentation detection can be conducted; however structural information on DTs would not be leveraged. Also, features of argumentation can potentially be measured in terms of maximal common sub-DTs, but such nearest neighbor learning is computationally intensive and too sensitive to errors in DT construction [Galitsky et al., 2015].

Therefore a CDT-kernel learning approach is selected which applies SVM learning ([Joty and Moschitti 2014], [Wang et al., 2010]) to the feature space of all sub-CDTs of the CDT for a given text where a heated argument is being detected.

We combined Stanford NLP parsing, coreferences, entity extraction, DT construction (discourse parser, [Surdeanu et al., 2016] and [Joty et al., 2013]), VerbNet and Tree Kernel builder into one system available at https://github.com/bgalitsky/relevance-based-on-parse-trees.

Our second example is a regular discourse tree for a text advising on how to behave communicating a heated argument [Inspiyr 2018].

> When you are in the middle of an argument, it can be easy to get caught up in the heat of the moment and say something that makes the situation even worse. Nothing can make someone more frenzied and hysterical than telling them to calm down. It causes the other person to feel as if you are putting the blame for the elevation of the situation on them. Rather than actually helping them calm down, it comes off as patronizing and will most likely make them even angrier.

A default DT for text such as a work of fiction or a scientific article, introducing and explaining a subject, would have default rhetoric relation of elaboration (as well as joint, attribution, background). This DT in addition has topic-comment on the top level, and also enablement, which indicates a peculiar logic flow. We will apply machine learning approach with extensive dataset of DT examples to observe a specific features of DTs associated with heated argumentation. In our earlier studies [Galitsky et al 2018] we developed a technique to extract and learn logical argumentation from text, and now we will apply it to heated arguments.

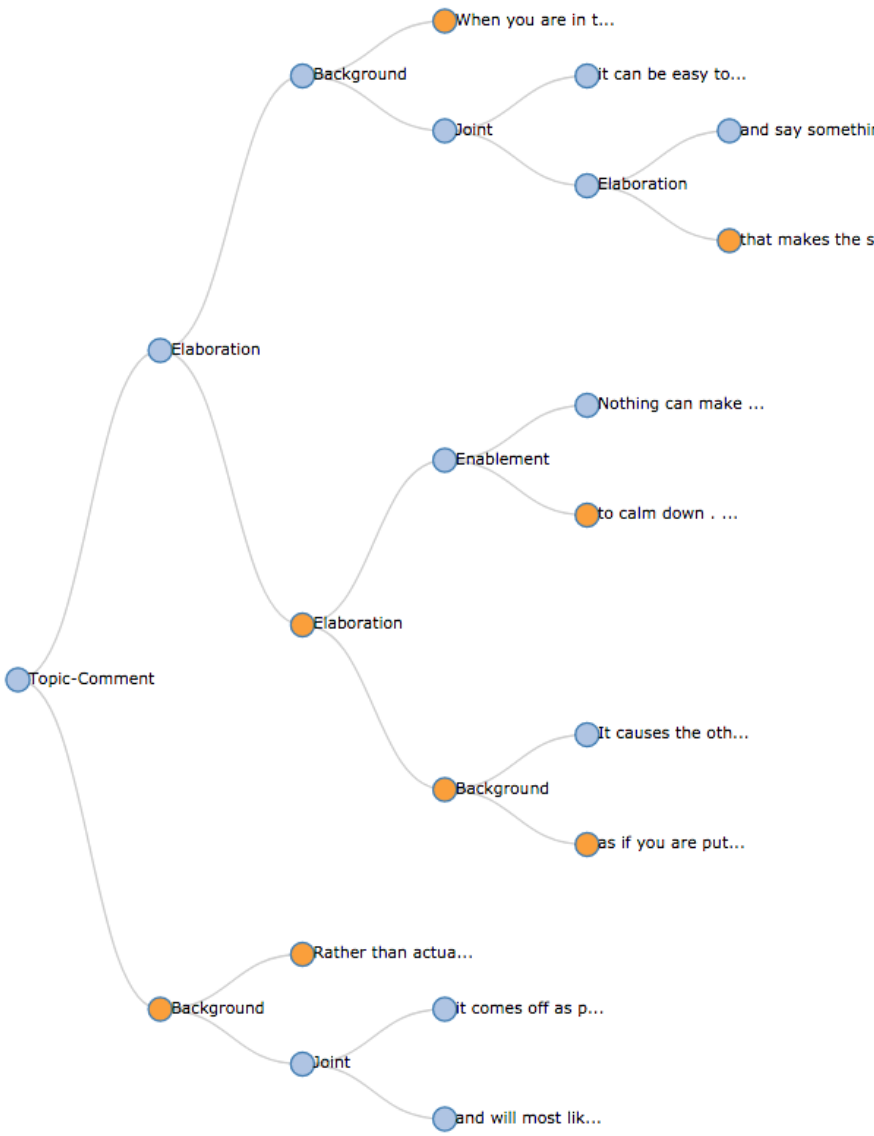


Fig. 2: The DT for a text advising on how to behave communicating an argument

3. Assessing Validity of Extracted Argument Patterns via Dialectical Analysis

To convince an addressee, a message needs to include an argument and its structure needs to be valid. Once an argumentation structure extracted from text is represented via CDT, we need to verify that the main point (target claim) communicated by the author is not logically attacked by her other claims. For a given domain, this claim is known (such as innocent or guilty, winning or loosing case, complaint is valid or not, violation has occurred or not). Most facts and clauses are pre-specified in a vertical domain ontology (the static part) and some of them are extracted from text via CDT (those can be less reliable).

To assess the validity of the argumentation, Defeasible Logic Programming (DeLP) approach is selected, an argumentative framework based on logic programming ([García and Simari 2004], [Alsinet et al 2008]), and present an overview of the main concepts associated with it.

A DeLP is a set of facts, strict rules P of the form (A:-B), and a set of defeasible rules D of the form A-<B, whose intended meaning is "if B is the case, then usually A is also the case". Let P=(P, D) be a DeLP program and L a ground literal.

Defeasible Rules Prepared In Advance
rent_receipt -< rent_deposit_transaction.
rent_deposit_transaction -< contact_tenant.
¬rent_deposit_transaction -<contact_tenant, three_days_notice_is_issued.
¬rent_deposit_transaction -< rent_is_overdue.
¬repair_is_done -< rent_refused, repair_is_done. repair_is_done -<
rent_is_requested.
¬rent_deposit_transaction -< tenant_short_on_money, repair_is_done.
¬repair_is_done -< repair_is_requested.
¬repair_is_done -<rent_is_requested.
¬repair_is_requested -< stay_unrepaired.
¬repair_is_done -< stay_unrepaired.
Target Claim to be Assessed
?—rent_receipt
Clauses Extracted from text
repair_is_done -< rent_refused.
Facts from text
contact_tenant. rent_is_requested. rent_refused. remind_about_repair.
three_days_notice_is_issued.
rent_is_overdue. stay_unrepaired.

Fig. 3a: An example of a Defeasible Logic Program for heated argument extracted from text

Let us now build an example of a DeLP for legal reasoning about facts extracted from text (Fig. 3). A judge hears an eviction case and wants to make a judgment on whether rent was provably paid (deposited) or not (denoted as *rent_receipt).*

An input is a text where a defendant is expressing his point. Underlined words form the clause in DeLP, and the other expressions formed the facts (Fig. 3b).

> *The landlord contacted me, the tenant, and the rent was requested. However, I refused the rent since I demanded repair to be done. I reminded the landlord about necessary repairs, but the landlord issued the tree-day notice confirming that the rent was overdue. Regretfully, the property still stayed unrepaired.*

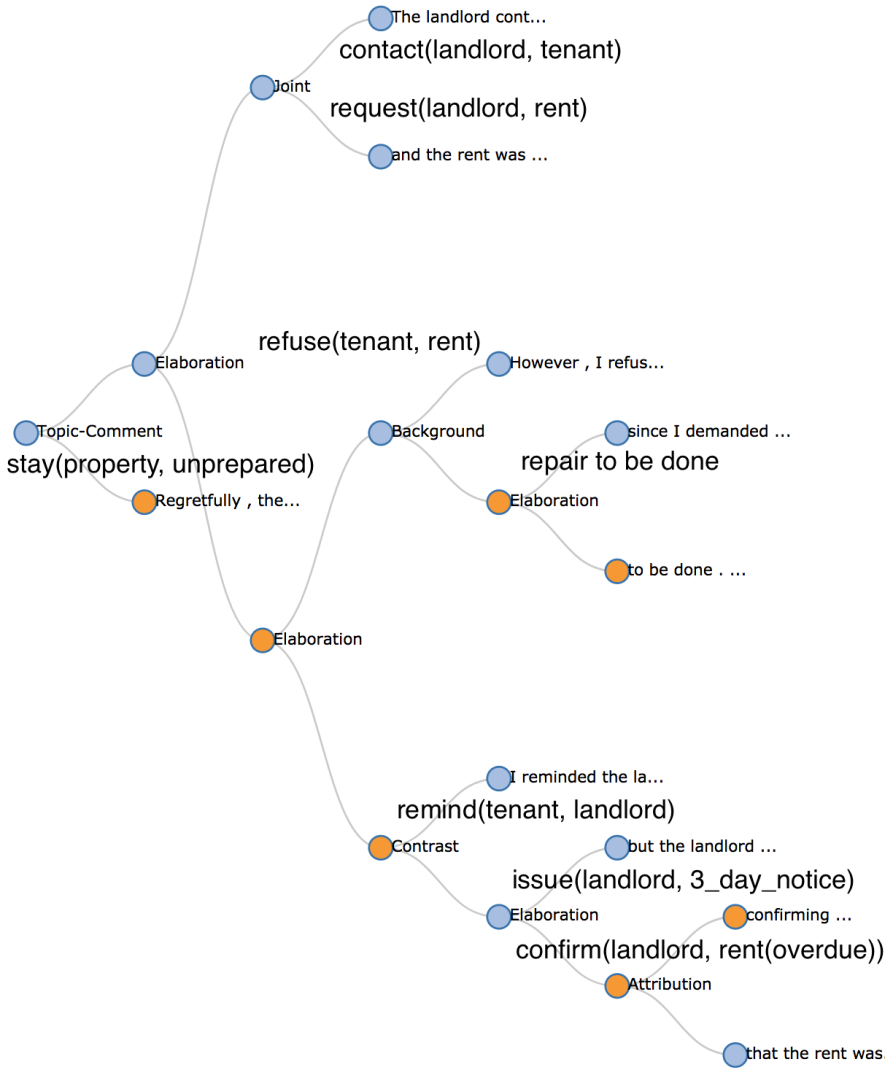


Fig 3b: Text of a complaint and its CDT

A *defeasible derivation* of L from P consists of a finite sequence $L_1, L_2, \ldots, L_n = L$ of ground literals, such that each literal L_i is in the sequence because:

(a) L_i is a fact in Π, or
(b) there exists a rule R_i in P (strict or defeasible) with head L_i and body $B_1, B_2, \ldots, B_k$ and every literal of the body is an element L_j of the sequence appearing before L_j ($j < i$).

Let h be a literal, and P=(Π, Δ) a DeLP program. We say that <A, h> is an *argument* for h, if A is a set of defeasible rules of Δ, such that:

1. there exists a defeasible derivation for h from =(Π ∪ A);
2. the set (Π ∪ A) is non-contradictory; and
3. A is minimal: there is no proper subset A_0 of A such that A_0 satisfies conditions (1) and (2).

Hence an argument <A, h> is a minimal non-contradictory set of defeasible rules, obtained from a defeasible derivation for a given literal h associated with a program P.

We say that $<A_1, h_1>$ *attacks* $<A_2, h_2>$ iff there exists a sub-argument <A, h> of $<A_2, h_2>$ ($A \subseteq A_1$) such that h and h_1 are inconsistent (i.e. $\Pi \cup \{h, h_1\}$ derives complementary literals). We will say that $<A_1, h_1>$ *defeats* $<A_2, h_2>$ if $<A_1, h_1>$ attacks $<A_2, h_2>$ at a sub-argument <A, h> and $<A_1, h_1>$ is strictly preferred (or not comparable to) <A, h>. In the first case we will refer to $<A_1, h_1>$ as a *proper defeater*, whereas in the second case it will be a *blocking defeater*. Defeaters are arguments which can be in their turn attacked by other arguments, as is the case in a human dialogue. An *argumentation line* is a sequence of arguments where each element in a sequence defeats its predecessor. In the case of DeLP, there are a number of *acceptability* requirements for argumentation lines in order to avoid fallacies (such as circular reasoning by repeating the same argument twice).

Target claims can be considered DeLP queries which are solved in terms of dialectical trees, which subsumes all possible argumentation lines for a given query. The definition of dialectical tree provides us with an algorithmic view for discovering implicit self-attack relations in users' claims. Let $<A_0, h_0>$ be an argument (target claim) from a program P. A *dialectical tree* for $<A_0, h_0>$ is defined as follows:

1. The root of the tree is labeled with $<A_0, h_0>$
2. Let N be a non-root vertex of the tree labeled $<A_n, h_n>$ and $\Lambda = [<A_0, h_0>, <A_1, h_1>, \ldots, <A_n, h_n>]$ (the sequence of labels of the path from the root to N). Let $[<B_0, q_0>, <B_1, q_1>, \ldots, <B_k, q_k>]$ all attack $<A_n, h_n>$. For each attacker $<B_i, q_i>$ with acceptable argumentation line $[\Lambda, <B_i, q_i>]$, we have an arc between N and its *child* N_i.

A labeling on the dialectical tree can be then performed as follows:

1. All leaves are to be labeled as U-nodes (undefeated nodes).
2. Any inner node is to be labeled as U-node whenever all its associated children nodes are labeled as D-nodes.
3. Any inner node is to be labeled as D-node whenever at least one of its associated children nodes is labeled as U-node.

After performing this labeling, if the root node of the tree is labeled as a U-node, the original argument at issue (and its conclusion) can be assumed as *justified* or *warranted*.

In our DeLP example, the literal *rent_receipt* is supported by

$<A, rent\_receipt> = <\{ (rent\_receipt -< rent\_deposit\_transaction), (rent\_deposit\_transaction -< tenant\_short\_on\_money)\}, rent\_receipt>$ and there exist three defeaters for it with three respective argumentation lines: $<B_1, \neg rent\_deposit\_transaction> = <\{(\neg rent\_deposit\_transaction -< tenant\_short\_on\_money, three\_days\_notice\_is\_issued)\}, rent\_deposit\_transaction>$.

$<B_2, \neg rent\_deposit\_transaction> = <\{(\neg\ rent\_deposit\_transaction -< tenant\_short\_on\_money, repair\_is\_done), (repair\_is\_done -< rent\_refused)\}, rent\_deposit\_transaction>$.

$<B_3, \neg rent\_deposit\_transaction> = <\{(\neg rent\_deposit\_transaction -< rent\_is\_overdue)\}, rent\_deposit\_transaction>$.

The first two are proper defeaters and the last one is a blocking defeater. Observe that the first argument structure has the counter-argument, $<\{rent\_deposit\_transaction -< tenant\_short\_on\_money\}, rent\_deposit\_transaction)$, but it is not a defeater because the former is more specific. Thus, no defeaters exist and the argumentation line ends there. B_3 above has a blocking defeater

$<\{(rent\_deposit\_transaction -< tenant\_short\_on\_money)\}, rent\_deposit\_transaction>$

which is a disagreement sub-argument of $<A, rent\_receipt>$ and it cannot be introduced since it gives rise to an unacceptable argumentation line. B_2 has two defeaters which can be introduced:

$<C_1, \neg repair\_is\_done >$, where $C_1 = \{(\neg repair\_is\_done -< rent\_refused, repair\_is\_done), (repair\_is\_done -< rent\_is\_requsted)\}$, a proper defeater, and $<C_2, \neg repair\_is\_done >$,

where $C_2=\{(\neg repair\_is\_done -< repair\_is\_requested)\}$ is a blocking defeater. Hence one of these lines is further split into two; C_1 has a blocking defeater that can be introduced in the line

$<D_1, \neg repair\_is\_done >$, where $D_1= <\{(\neg repair\_is\_done -< stay\_unrepaired)\}$.

D_1 and C_2 have a blocking defeater, but they cannot be introduced, because they make the argumentation line inacceptable. Hence the state *rent_receipt* cannot be reached, as the argument supporting the literal *rent_receipt* is not warranted. The dialectical tree for *A* is shown in Fig. 4.

Having shown how to build dialectic tree, we now ready to outline the algorithm for validation the domain-specific claim for arguments extracted from text:

1. Build a DT from input text;
2. Attach communicative actions to its edges to form CDT;
3. Extract subjects of communicative actions attached to CDT and add to 'Facts' section;

4. Extract the arguments for rhetoric relation *contrast* and communicative actions of the class *disagree* and add to 'Clauses Extracted FromText' section;
5. Add domain-specific section to DeLP;
6. Having the DeLP formed, build a dialectical tree and assess the claim.

We used [Tweety 2017] system for DeLP implementation. The Tweety package contains several classes for dealing with abstract argumentation frameworks which can be imported programmatically using specific methods. Tweety supports reasoning relying on the extension-based approaches of grounded, stable, complete, preferred, ideal, semistable, CF2, and stage semantics as well as the ranking-based approaches of [Grossi & Modgil, 2015].

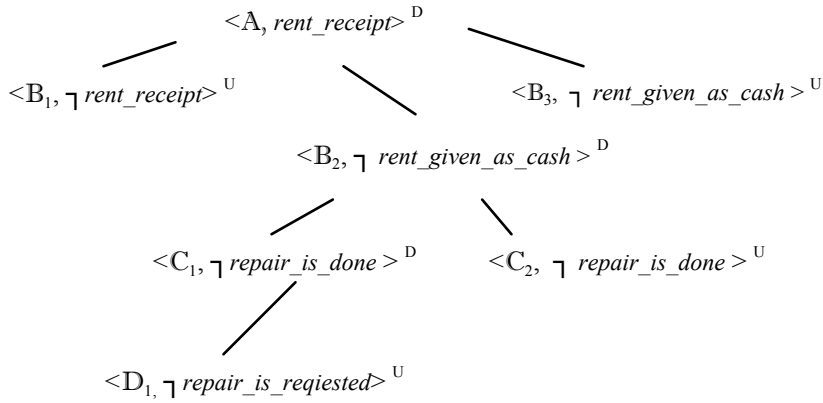


Fig. 4: Dialectical tree for target claim rent_receipt

4. Evaluation of Detection and Validation of Affective Arguments

The objective of argument detection task is to identify all kinds of arguments, not only ones associated with customer complaints. We formed the *positive* dataset from textual customer complaints dataset (Galitsky et al., 2009, and https://github.com/bgalitsky/relevance-based-on-parse-trees/blob/master/src/test/resources/opinions-FinanceTagged.xls.zip, scraped from consumer advocacy site PlanetFeedback.com. This dataset is used for both argument detection and argument validity tasks.

Table 1: Evaluation results for argument detection

Method / sources	P	R	F1
Bag-of-words	57.2	53.1	55.07
WEKA-Naïve Bayes	59.4	55.0	57.12
SVM TK for RST and CA (full parse trees)	77.2	74.4	75.77
SVM TK for DT	63.6	62.8	63.20
SVM TK for CDT	82.4	77.0	79.61

For the *negative* dataset, only for the affective argument detection task, we used Wikipedia, factual news sources, and also the component of [Lee, 2001] dataset that includes such sections of the corpus as: ['tells'], instructions for how to use software; ['tele'], instructions for how to use hardware", and [news], a presentation of a news article in an objective, independent manner, and others. Further details on the data set are available in [Galitsky et al 2015].

A baseline approach relies on keywords and syntactic features to detect argumentation (Table 1). Frequently, a coordinated pair of communicative actions (so that at least one has a negative sentiment polarity related to an opponent) is a hint that logical argumentation is present. This naïve approach is outperformed by the top performing TK learning CDT approach by 29%. SVM TK of CDT outperforms SVM TK for RST+CA and RST + full parse trees [Galitsky 2017] by about 5% due to noisy syntactic data which is frequently redundant for argumentation detection.

SVM TK approach provides acceptable F-measure but does not help to explain how exactly the affective argument identification problem is solved, providing only final scoring and class labels. Nearest neighbor maximal common sub-graph algorithm is much more fruitful in this respect [Galitsky et al 2015]. Comparing the bottom two rows, we observe that it is possible, but infrequent to express an affective argument without CAs.

Assessing logical arguments extracted from text, we were interested in cases where an author provides invalid, inconsistent, self-contradicting cases. That is important for CRM systems focused on customer retention and facilitating communication with customer [Galitsky et al 2009]. The domain of residential real estate complains was selected and DeLP thesaurus was built for this domain. Automated complaint processing system is essential, for example, for property management companies in their decision support procedures [Constantinos et al 2003].

Table 2: Evaluation results for argument validation

Types of complaints	P	R	F1 of validation	F1 of total
Single rhetoric relation of type *contrast*	87.3	15.6	26.5	18.7
Single communicative action of type *disagree*	85.2	18.4	30.3	24.8
Two or three specific relations or communicative actions	80.2	20.6	32.8	25.4
Four and above specific relations or communicative actions	86.3	16.5	27.7	21.7

In our validity assessment we focus on target features related to how a given complaint needs to be handled, such as *compensation_required, proceed_with_eviction, rent_receipt* and others.

Complaint validity assessment results are shown in Table 2. In the first and second rows, we show the results of the simplest complaint with a single rhetoric relation such as *contrast* and a single CA indicating an extracted argumentation attack

relation respectively. In the third row we assess complaints of average complexity, and in the bottom row—most complex, longer complaints in terms of their CDT. The third column shows detection accuracy for invalid argumentation in complaints in a standalone argument validation system. Finally, the fourth column shows the accuracy of the integrated argumentation extraction and validation system.

For decision support systems, it is important to maintain a low false positive rate. It is acceptable to miss invalid complaints, but for a detected invalid complain, confidence should be rather high. If a human agent is recommended to look at a given complaint as invalid, her expectations should be met most of the times. Although F1 measure of the overall argument detection and validation system is low in comparison with modern recognition systems, it is still believed to be usable as a component of a CRM decision support system.

5. Conclusions

We observed that relying on discourse tree data, one can reliably detect patterns of affective argumentation. Communicative discourse trees then become a source of information to form a defeasible logic program to validate an argumentation structure. Although the performance of the former being about 80% is significantly above that of the latter (29%), the overall pipeline can be useful for detecting cases of invalid heated argumentation, which are important in decision support for CRM. Once it is possible to extract amplified, heated arguments, in our future studies we will proceed to combining mining and reasoning about general arguments, not necessarily accented by a sentiment.

We anticipate the difficulties in adopting the argumentation pipeline in industry. Today, sentiment analysis is extensively used by companies to understand which features of products and services are appreciated by customers and which need improvement. Deeper understanding of customer complaints, implemented in this study, would reveal shady corporate practices and would put a blame on certain company management individuals responsible for respective product limitations and customer support deficiencies. Internal corporate policies and internal conflicts of interest between management structures could potentially be affected by findings produced by the argumentation pipeline, and a significant number of corporate management members might oppose obtaining these findings. Hence a series of issues outside of the technology area might prevent argumentation pipeline from being deployed in industry. [Galitsky 2016] addressed the corporate conflict of interest models from the standpoint of multiagent systems; the results show that a corporate multiagent system can involve into behavioral forms distant from being rational or competent.

In this paper we attempted to combine the best of both worlds, argumentation mining from text and reasoning about the extracted argument. Whereas applications of either technology are limited, the whole argumentation pipeline is expected to find a broad range of applications. In this work we focused on a very specific legal area such as customer complaints, but it is easy to see a decision support system employing the proposed argumentation pipeline in other domains of CRM.

An important finding of this study is that argumentation structure can be discovered via the features of extended discourse representation, combining information on how an author organizes her thoughts with information on how involved agents communicate these thoughts. Once a communicative discourse tree is formed and identified as being correlated to argumentation, a defeasible logic program can be built from this tree and the dialectical analysis can validate the main claim.

References

1. *Mann, William and Sandra Thompson* (1988). Rhetorical structure theory: Towards a functional theory of text organization. Text-Interdisciplinary Journal for the Study of Discourse, 8(3):243–281.
2. *Alsinet, T., Carlos Iván Chesñevar, Lluis Godo, Guillermo Ricardo Simari* (2008). A logic programming framework for possibilistic argumentation: Formalization and logical properties. Fuzzy Sets and Systems 159(10): 1208–1228
3. *Mikolov, Tomas, Chen, Kai, Corrado; G.S., Dean; Jeffrey* (2015). Computing numeric representations of words in a high-dimensional space. US Patent 9,037,464, Google, Inc.
4. *Wang, W., Su, J., Tan, C. L.* (2010). Kernel Based Discourse Relation Recognition with Temporal Ordering Information. In *ACL*.
5. *Wei Feng, Vanessa and Graeme Hirst* (2014). A linear-time bottom-up discourse parser with constraints and post-editing. In ACL.
6. *Joty, Shafiq R, Giuseppe Carenini, Raymond T Ng, and Yashar Mehdad* (2013). Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *ACL (1)*, pages 486–496.
7. *Joty, Shafiq R and A. Moschitti* (2014). Discriminative Reranking of Discourse Parses Using Tree Kernels. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).
8. *Kipper, K. Korhonen, A., Ryant, N. and Palmer, M.* (2008). A large-scale classification of English verbs. Language Resources and Evaluation Journal, 42, pp. 21–40.
9. *Surdeanu, Mihai, Thomas Hicks, and Marco A. Valenzuela-Escarcega* (2015) Two Practical Rhetorical Structure Theory Parsers. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics—Human Language Technologies: Software Demonstrations (NAACL HLT).
10. *Galitsky, B., MP González, CI Chesñevar* (2009). A novel approach for classifying customer complaints through graphs similarities in argumentative dialogue. Decision Support Systems, 46–3, 717–729.
11. *Galitsky, B.* (2013). Machine learning of syntactic parse trees for search and classification of text. *Engineering Application of AI* , 26(3) 1072–91.
12. *Galitsky, B.* (2015). Detecting Rumor and Disinformation by Web Mining. AAAI Spring Symposium, 2015.
13. *Galitsky, B., Ilvovsky, D. and Kuznetsov S. O.* (2015). Rhetoric Map of an Answer to Compound Queries. ACL-2, 681–686.
14. *Galitsky, B.* (2016). Computational Autism. Springer, London UK.

15. *Galitsky, B.* (2017a). Matching parse thickets for open domain question answering, Data & Knowledge Engineering, Volume 107, January 2017, Pages 24–50.
16. *Galitsky, B.* (2017b). Using Extended Tree Kernel to Recognize Metalanguage in Text. In Uncertainty Modeling, Volume 683 of the series Studies in Computational Intelligence pp. 71–96, Springer.
17. *Constantinos J. Stefanou, Christos Sarmaniotis, Amalia Stafyla.* (2003). CRM and customer-centric knowledge management: an empirical research, Business Process Management Journal, Vol. 9 Issue: 5, pp.617–634.
18. *Tweety* (2016). Last downloaded Dec 12, 2016. https://javalibs.com/artifact/net.sf.tweety.arg/delp.
19. *Thimm, M.* (2014) Tweety—A Comprehensive Collection of Java Libraries for Logical Aspects of Artificial Intelligence and Knowledge Representation. In Proceedings of the 14th International Conference on Principles of Knowledge Representation and Reasoning (KR'14). Vienna, July, 2014
20. *Garcia, Alejandro and Guillermo R. Simari.* (2004) Defeasible Logic Programming: An Argumentative Approach. Theory and Practice of Logic Programming 4(1–2):95–138, 2004.
21. *Amgoud, Leila, Philippe Besnard, Anthony Hunter.* (2015). Representing and Reasoning About Arguments Mined from Texts and Dialogues. ECSQARU 2015: 60–71.
22. *Bondarenko, A., Dung, P., Kowalski, R., Toni, F.* (1997) An abstract, argumentation-theoretic approach to default reasoning. Artificial Intelligence 93, 63–101.
23. *Inspiyr* (2018). 5 Things You Should Never Say During A Heated Argument. https://inspiyr.com/heated-argument/.
24. *Vishnu S. Pendyala, Silvia Figueira* (2015) Towards a truthful world wide web from a humanitarian perspective. Global Humanitarian Technology Conference (GHTC), 2015 IEEE, Issue Date: 8–11 Oct. 2015.
25. *Grossi, D., Modgil, S.* (2015) On the graded acceptability of arguments. In: Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI'15). pp. 868–874.
26. *Toni, F.* (2014). A tutorial on assumption-based argumentation. Argument & Computation 5(1), 89–117.
27. *Moens, Marie-Francine, Erik Boiy, Raquel Mochales Palau, and Chris Reed* (2007). Automatic detection of arguments in legal texts. In Proceedings of the 11th International Conference on Artificial Intelligence and Law, ICAIL '07, pages 225–230, Stanford, CA, USA.
28. *Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis* (2015). Argument extraction from news. In Proceedings of the 2nd Workshop on Argumentation Mining, pages 56–66, Denver, CO, USA.
29. *Stab, C. and Gurevych, I.* (2016) Recognizing the absence of opposing arguments in persuasive essays. ACL 2016.
30. *Walton, D., Reed, C., Macagno, F.* (2008) Argumentation Schemes. Cambridge University Press. .
31. *Pisarevskaya, Dina, Tatiana Litvinova, Olga Litvinova* (2017) Deception Detection for the Russian Language: Lexical and Syntactic Parameters. Proceedings of the 1st Workshop on Natural Language Processing and Information Retrieval / RANLP.

ВЛИЯНИЕ СИНТАКСИСА НА ПРОСОДИЮ: ДАННЫЕ ОДНОГО ЭКСПЕРИМЕНТА НАД РУССКИМ ПИСЬМЕННЫМ ТЕКСТОМ

Гращенков П. В. (pavel.gra@gmail.com)[1]
МГУ им. М. В. Ломоносова; ИВ РАН, МПГУ; Москва, Россия

Кириллова А. А. (anastasya_kirillova@hotmail.com),
Смирнова О. С. (kisaolga@mail.ru)
МГУ им. М. В. Ломоносова; Москва, Россия

THE INFLUENCE OF SYNTAX ON PROSODY: THE EXPERIMENTAL DATA FROM A STUDY OF ONE RUSSIAN TEXT

Grashchenkov P. V. (pavel.gra@gmail.com)
Lomonosov Moscow State University; IOS RAS,
Moscow Pedagogical State University; Moscow, Russia

Kirillova A. A. (anastasya_kirillova@hotmail.com),
Smirnova O. S. (kisaolga@mail.ru)
Lomonosov Moscow State University; Moscow, Russia

The paper examines dependencies between the syntactic and prosodic structure with particular attention to the pausation and different levels of prosodic boundary strength. The research is based on the prosodic data markup for a spoken Russian text and the manual tagging of this text with the relevant syntactic constituent boundaries. Two types of structures, the finite clause and the asyndetic coordination, exhibit a strong positive correlation with the appearance of a pause and the perceptual prosodic boundary. We also demonstrate the presence of a substantial correlation between the syntactic embedding depth and prosodic boundaries. The results of our research show a significant connection between some of the initially proposed syntactic factors and prosodic structure. We thus anticipate that prosodic modules of TTS systems can benefit from taking certain syntactic information into consideration.

Key words: syntactic structure, prosodic structure, Russian, pausation, prosodic boundary, correlation analysis

[1] Участие П. В. Гращенкова поддержано грантом РНФ № 16-18-02003, реализуемым в ФГБОУ ВО МПГУ.

1. Введение: синтаксическая и просодическая организация высказывания[2]

Предложения как законченные формы высказывания имеют определенную синтаксическую структуру. Одновременно с этим любое предложение может члениться на фонетические составляющие (фонетические фразы, синтагмы). Подобное членение вслед за [Кривнова 2015] и другими работами мы будем называть просодическим (ПЧ). Основные фонетические средства ПЧ — паузация и переломы тона, некоторое влияние, кроме того, оказывают и такие факторы как громкость, темп и фонация, см. [Кривнова 2015], [Кодзасов, Кривнова 2001: 304–330].

Просодические границы между синтагмами могут быть различной глубины. Слабее всего границы внутри фонетического слова и акцентной группы, следующие по силе границы возникают между фонетическими словами / акцентными группами внутри фонетической синтагмы, далее идут границы между фонетическими синтагмами / акцентными группами, и, наконец, между интонационными фразами и на границе смежных высказываний, см. [Кривнова 2015].

В [Sanderman 1996] на материале нидерландского языка впервые была исследована «объективная» природа ПЧ. Было установлено, что перцептивное восприятие разных носителей хорошо соотносится между собой — определение границ ПЧ разными носителями в целом совпадает. Сандерман также определила, что носители сходятся примерно на 5 уровнях глубины ПЧ, а наибольший вес по сравнению с другими фонетическими факторами имеет паузация.

Как показано в экспериментах, проведенных под руководством О. Ф. Кривновой, см. [Кривнова 2015], [Смирнова 2017], для русского языка также оказываются релевантны 5 уровней глубины ПЧ. Более того, как было установлено в рамках данных экспериментов, для русского языка 5 уровней перцептивно различаемых носителями просодических швов (ПШ) также значимо коррелируют с длительностью паузы. Швам разного уровня соответствуют паузы определенной длительности.

Существуют различные точки зрения относительно связи синтаксиса и ПЧ. Некоторые исследователи предполагают, что синтаксические иерархии и синтаксические составляющие являются основой для действия последующего просодического компонента, см. [Halliday 1967], [Selkirk 1978], [Янко 2008]. Альтернативная точка зрения состоит в том, что ПЧ не основывается на синтаксисе, а наблюдаемые случаи совпадения границ подчиняются определенным когнитивным механизмам, см. [Croft 1995], [Lahiri, Plank 2010].

В работе [Fach 1999] соответствие просодических и синтаксических границ для английского языка было оценено на уровне 65 %. Это показывает, что синтаксическая структура явно заслуживает внимания при определении границы ПЧ.

[2] Авторы выражают глубокую благодарность О. Ф. Кривновой за помощь в подготовке и проведении данного исследования. Авторы также признательны Оргкомитету конференции «Диалог» и анонимным рецензентам за сделанные наблюдения и замечания.

Главная задача, поставленная в данной работе, — исследовать вопрос о том, насколько синтаксическая структура влияет на просодическое членение в русском языке.

2. Паузация и синтаксическая структура

В данном разделе будут представлены результаты предпринятого нами экспериментального исследования зависимости паузации от синтаксической структуры.[3]

2.1. Прикладные и экспериментальные исследования влияния синтаксиса на просодическое членение

В [Кривнова, Чардин 1999] говорится как о необходимости учета синтаксической информации для постановки пауз, так и о технической сложности получения адекватных данных о синтаксической структуре. В работе [Лобанов 2008] предлагается элементарный алгоритм построения синтаксической структуры для русского языка. Данная синтаксическая структура используется для простановки просодических границ «напрямую», на основании ее также проставляется фразовое ударение. Позже в данную систему был интегрирован синтаксический анализатор лингвистического процессора ЭТАП-3, см. [Иомдин, Лобанов 2009], [Иомдин, Лобанов, Гецевич 2011].

Хорошие результаты по качеству простановки просодических границ для русского языка задекларированы в разработках ООО «ЦРТ», см. [Хомицевич, Соломенник 2010], [Рыбин, Чистиков, Хомицевич 2014]. Применяемый алгоритм использует правила и статистические модели и основывается на информации о пунктуации, количестве слов и слогов, грамматической форме и синтаксических связях. Как показывают авторы, работа над качеством синтаксических связей приводит к улучшению качества просодической разметки.

Необходимость учета синтаксической структуры при построении просодических границ признается большинством ученых. Есть, однако, две проблемы, преодолеть которые на данном этапе практически не удалось: i) отсутствие четкого понимания того, как именно синтаксическая структура влияет на просодическую организацию и ii) недостаточное качество информации о синтаксической структуре, предоставляемое системами автоматического анализа.

---

[3] Один из рецензентов задался двумя вопросами в связи с выбранной нами методикой анализа результатов дикторского произнесения письменного текста. Первый вопрос — почему не используется анализ спонтанной речи, второй — почему анализ пауз бинарный (да/нет). Относительно использования спонтанной речи — мы считаем (см. Заключение), что это необходимо сделать в будущем. Начать с письменного текста было решено по причине доступности для него синтаксической разметки (см. раздел 3) а также потому, что у проекта есть и прикладной аспект — задача установления факторов, релевантных при автоматическом озвучивании письменного текста. По второму вопросу: недостаток анализа бинарной оппозиции при паузации отчасти снимается анализом ПШ разной глубины (см. раздел 3). Бинарный анализ способен дать общие сведения о корреляции, которые впоследствии, безусловно, необходимо уточнить.

Тем не менее, даже возможное наличие ошибок и недостаточность точных знаний о том, как синтаксис влияет на ПШ в конкретном случае, не означают, что учет синтаксической информации бессмысленен, см., например, [Tepperman, Nava 2011]. В указанной работе на материале английского языка предложен алгоритм преобразования полных синтаксических деревьев в цепочки просодических ярлыков с их промежуточным представлением в виде иерархической структуры. По утверждению авторов, это позволило получить прирост в качестве относительно алгоритма, основанного только на n-граммах из цепочек частеречных ярлыков.

Попытка полагаться при определении границ ПЧ на синтаксическую информацию, даже если применяется наименее ресурсоемкий «грубый» парсинг, зачастую похожа на гадание. Причина этого — отсутствие четких представлений о том, как правила синтаксиса отображаются в правила ПЧ.

В данной работе ставится задача выявить синтаксические конструкции, оказывающие влияние на ПЧ. В случае их обнаружения информацию о них необходимо в дальнейшем учитывать в прикладных системах с одной стороны, а с другой — попытаться осмыслить факты корреляция между синтаксисом и просодией посредством теоретического аппарата лингвистики, см. возможные направления анализа в [Гращенков, Кириллова, Смирнова 2018: раздел 2.2].

2.2. Языковые данные

Исследование проводилось по тексту И. Грековой (Е. С. Вентцель) «Люда Величко» (часть повести «Кафедра»). Отрывок, содержащий 2712 словоформ, был озвучен диктором, после чего получил просодическую и синтаксическую разметку. Просодическая разметка содержит информацию о длительности пауз (всего их отмечено 737) и наличии просодических швов (градация — от 0 до 5).

Данные о паузации исходно представляли собой длину паузы в миллисекундах. При получении результатов они рассматривались как бинарный признак: при наличии паузы ненулевой длительности считалось, что пауза есть, а в случае, когда длительность межсловного интервала равнялось нулю, — что ее нет.

Чтобы получить свидетельства (отсутствия) связи синтаксической составляющей с просодическим оформлением было выделено несколько типов потенциально релевантных синтаксических структур. Далее была проведена ручная разметка текста маркерами таких составляющих. Ниже представлены результаты этого исследования.

2.3. Связь типа синтаксической проекции с паузацией

Большинство выделенных для анализа синтаксических составляющих, см. [Гращенков, Кириллова, Смирнова 2018: раздел 2.4] продемонстрировало наличие зависимости с паузацией согласно критерию независимости номинальных признаков χ^2 и «точному» критерию Фишера при уровне значимости (вероятности ошибки 1-го рода) 0,05. Отсутствие корреляций между ожидаемым оформлением (=паузацией) было отмечено лишь для адъективных оборотов, инверсии и нулевых связок (по обоим статистическим критериям).

Ниже приведено количество оборотов каждого типа с данными по средней длительности пауз.

a) финитная клауза (CP) в абсолютном конце предложения

Можно утверждать, что граница финитной клаузы на конце предложения всегда связана с паузацией.

Таблица 1. Данные о паузах на границах конечных финитных клауз

паузация	средняя длит. паузы (мс)
235/235 (100%)	893

b) финитная клауза (CP) внутри сложного предложения

Финитная клауза в составе сложного предложения также соответствует паузе.

Таблица 2. Данные о паузах на границах внутренних финитных клауз

паузация	средняя длит. паузы (мс)
217/263 (83%)	376

c) деепричастный оборот

Левая и правая границы деепричастного оборота также связаны с паузами.

Таблица 3. Данные о паузах на границах деепричастных оборотов

	паузация	средняя длит. паузы (мс)
левая граница	11/16 (69%)	182
правая граница	5/6 (83%)	356
в целом	16/22 (73%)	236

d) сочинение без союза

О значимом влиянии бессоюзного сочинения на паузацию можно говорить во всех случаях, даже если размер выборки небольшой (например, в случае PP — согласно критерию Фишера $p < 0{,}05$).

Таблица 4. Данные о паузах при бессоюзном сочинении

	паузация	средняя длит. паузы (мс)
CP	107/120 (89%)	337
VP	40/57 (70%)	262
DP	36/50 (72%)	170
AP	15/21 (71%)	246

	паузация	средняя длит. паузы (мс)
PP	5/8 (63%)	124
в целом	203/258 (79%)	281

e) сочинение с союзом

Союзное сочинение в большинстве случаев соответствует паузам. Согласно критерию χ^2 и точному критерию Фишера зависимость значима ($p << 0.05$).

Таблица 5. Данные о паузах при союзном сочинении (типы составляющих)

	паузация	средняя длит. паузы (мс)
CP	25/30 (83%)	312
VP	22/26 (85%)	242
DP	9/15 (60%)	206
в целом	61/78 (78%)	259

f) причастный оборот

Причастный оборот связан с паузацией.

Таблица 6. Данные о паузах на границах причастных оборотов

	паузация	средняя длит. паузы (мс)
левая граница	10/15 (67%)	2000
правая граница	5/6 (83%)	212
в целом	15/21 (71%)	204

g) адъективный оборот

Данных по адъективным оборотам достаточно немного. Из них, однако, следует, что они не оказывают статистически значимого влияния на паузацию согласно точному критерию Фишера:

Таблица 7. Данные о паузах на границах адъективных оборотов

	паузация	средняя длит. паузы (мс)
левая граница	1/5 (20%)	25
правая граница	3/4 (75%)	62
в целом	4/9 (44%)	53

h) субстантивный оборот

Субстантивный оборот соответствует паузам при озвучивании текста.

Таблица 8. Данные о паузах на границах субстантивных оборотов

	паузация	средняя длит. паузы (мс)
левая граница	11/12 (92%)	293
правая граница	4/5 (80%)	203
в целом	15/17 (88%)	262

i) инверсия

Согласно обоим статистическим критериям, отмеченные случаи инверсии не оказывают значимого влияния на паузацию.

Таблица 9. Данные о паузах при инверсии

паузация	средняя длит. паузы (мс)
10/36 (28%)	176

j) подъем объекта или адъюнкта

Подъем синтаксической составляющей положительно связан с паузацией согласно критерию χ^2 ($p < 0{,}05$).

Таблица 10. Данные о паузах при подъеме составляющей

паузация	средняя длит. паузы (мс)
13/79 (16%)	165

k) нулевая связка

Примеры с нулевыми связками не продемонстрировали непосредственной связи с наличием паузы ни по одному из двух статистических критериев.

Таблица 11. Данные о паузах при нулевых связках

паузация	средняя длит. паузы (мс)
9/37 (24%)	174

l) эллипсис

Согласно критерию Фишера эллипсис влияет на паузацию ($p < 0{,}05$), отметим, однако, небольшой объем выборки.

Таблица 12. Данные о паузах при эллипсисе

паузация	средняя длит. паузы (мс)
1/19 (5%)	31

2.4. Связь типа синтаксической проекции с паузацией: итоги

Однозначные корреляции с паузой наблюдаются в случаях: a) финитной клаузы в абсолютном конце предложения, b) финитной клаузы внутри сложного предложения, c) деепричастного оборота, d) сочинения без союза, e) сочинения с союзом, f) причастного оборота, h) субстантивного оборота, j) подъема объекта или адъюнкта.

В случае эллипсиса, пункт l), корреляция между синтаксической конструкцией и наличием паузы также скорее имеет место.

Об отсутствии корреляции можно говорить в случаях: g) адъективного оборота, i) инверсии, k) нулевой связки.

Наличие запятой не всегда связано с паузацией. Это видно на примере адъективных оборотов, выделяющихся запятой, но не демонстрирующих связи с паузами. Возможно, таким образом, говорить о том, что синтаксические факторы влияют на паузацию не только опосредованно, т. е. через запятые, но также должны учитываться и независимо от запятых.

3. Некоторые корреляции между синтаксисом и ПЧ

Нами также было проведено пилотное исследование корреляции синтаксиса и границ ПЧ[4]. Перцептивные границы ПЧ были выделены на основании согласования данных пяти носителей (из двадцати принимавших участие в эксперименте).

3.1. Корреляции между синтаксическими факторами и ПШ

При рассмотрении ПЧ в контексте синтаксиса важно понимать, какое влияние тот или иной фактор оказывает на появление пауз в речи и насколько оно значимо. В связи с этим был предпринят корреляционный анализ выборки, полученной после обработки исходного озвученного текстового фрагмента. Ниже мы абстрагируемся от дробного деления синтаксических конструкций.

Для оценки степени зависимости между ПШ и синтаксическими конструкциями для конструкции каждого типа были посчитаны значения двух коэффициентов корреляции — Пирсона (линейного) и Спирмена (рангового). Так как на основе имеющихся данных нельзя исключить нелинейные корреляционные зависимости, использование рангового критерия представляется необходимым. В Таблице 13 приведены значения коэффициентов для факторов, для которых с вероятностью 95 % можно утверждать о наличии значимой взаимосвязи с ПШ. О выраженной корреляции можно говорить лишь в случае финитной клаузы и бессоюзного сочинения, в остальных случаях зависимость слабая (о сильной связи свидетельствуют значения коэффициентов, близкие к ±1, о слабой — близкие к 0).

---

[4] Авторы признательны О. Ф. Кривновой, С. В. Князеву, Е. В. Моисеевой и Л. М. Захарову за предоставленные данные по разметке.

Таблица 13. Результаты парного корреляционного анализа связи ПШ и синт. факторов

фактор	коэф. Пирсона	коэф. Спирмена
финитная клауза	**0,76**	**0,72**
бессоюзное сочинение	**0,35**	**0,42**
союзное сочинение	0,14	0,19
адъективный + причастный обороты	0,07	0,11
деепричастный оборот	0,07	0,10
субстантивный оборот	0,07	0,09

В процессе исследования была отдельно проанализирована связь маркеров синтаксических составляющих, полученных в результате ручной разметки, и глубины ПШ. В таблицах сопряжённых признаков 14 и 15 представлены распределения 2712 словоразделов исследуемого текста в зависимости от присутствия финитной клаузы и бессоюзного сочинения как факторов, наиболее явно коррелирующих с паузацией. Сила брейка ПШ принимает значение от 0 до 5.

Таблица 14. Распределение данных по уровням ПШ и наличию маркера финитной клаузы

финитная клауза	глубина ПШ						
	0	1	2	3	4	5	всего
0	**1808**	158	229	19	0	0	2214
1	19 (4%)	27	**193 (39%)**	**180 (36%)**	59	20	498
всего	1827	185	422	199	59	20	2712

Таблица 15. Распределение данных по уровням ПШ и наличию маркера бессоюзного сочинения

бессоюзное сочинение	глубина ПШ						
	0	1	2	3	4	5	всего
0	**1822**	138	263	154	55	20	2452
1	5 (2%)	47	**159 (61%)**	45	4	0	260
всего	1827	185	422	199	59	20	2712

Одновременно с меткой финитной клаузы наиболее вероятны ПШ уровня 2 и 3 (словоразделы с такой силой брейка составляют соответственно 39% и 36% всех маркеров границ клаузы). В случае маркера бессоюзного сочинения в 61% случаев глубина просодического шва равна 2. Данные таблиц сопряжённости, с одной стороны, еще раз подтверждают наличие связи между рассматриваемыми синтаксическими конструкциями и ПШ — лишь в 4% и 2% случаев

соответственно сила брейка при наличии фактора равна 0; коэффициент квадратичной связи Крамера V[5] оказался равен 0.77 для финитной клаузы и 0.5 для бессоюзного сочинения. С другой стороны, подобного рода информация дает возможность с определенной долей уверенности предсказывать глубину ПШ и длительность паузы для словоразделов с известным синтаксическим окружением. Подобного рода информацию полезно учитывать при работе над системами автоматического синтеза речи для получения наиболее естественной паузации.

3.2. Корреляции между глубиной синтаксического вложения и ПШ на основании разметки ЭТАП-3

В процессе анализа синтаксической информации были отдельно рассмотрены данные о границах составляющих, полученные из синтаксической разметки того же текста повести «Кафедра» Грековой, осуществленной в рамках проекта ЭТАП-3, см. [Дьяченко и др. 2015][6]. Благодаря разметке границ непосредственных составляющих появилась возможность изучения связи просодии и глубины вложения, соответствующей количеству скобок в структуре составляющих.

Гипотеза о зависимости просодической структуры от степени вложенности соответствующих синтаксических единиц согласуется с подходом в генеративной фонологии, известным как Правило Ядерного Ударения (Nuclear Stress Rule) и постулирующим обусловленность места акцента структурой составляющих. В упрощенном виде общепринятую на данный момент формулировку (см., например, [Cinque 1993]) можно представить так: фразовое ударение при озвучивании получает наиболее глубоко вложенная составляющая. ПЧ в сильной степени связано с фразовым ударением, следовательно, гипотеза о зависимости ПЧ от глубины синтаксического вложения также должна быть изучена.

В Таблице 16 приведены результаты корреляционного анализа для различных участков скобочной разметки на словоразделе и ПШ.

Таблица 16. Коэффициенты корреляции Пирсона между элементами скобочной разметки и ПШ

признак	коэффициент корреляции
число закрывающих скобок]	0,59
число открывающих скобок [	0,43

[5] Коэффициент V принимает значения между 0 и 1 и является нормированным значением статистики Пирсона (критерий χ^2 для проверки гипотезы независимости в таблицах сопряженных признаков). Близость его к нулю говорит о независимости, а в случае близости к 1 гипотеза независимости может быть отвергнута с малой вероятностью ошибки.

[6] Авторы статьи выражают благодарность коллегам из проекта ЭТАП-3 за предоставление корпуса с синтаксической разметкой СинТагРус. Авторы также благодарны О. М. Аншакову за помощь в преобразовании дерева зависимостей в структуру составляющих.

Можно видеть, что между количеством скобок и наличием просодического шва присутствует значимая связь, наиболее показательным по степени влияния на ПШ фактором оказывается число закрывающих скобок.

3.3. Корреляции между синтаксисом и ПЧ: результаты

Подытоживая материал данного раздела, сформулируем следующие результаты: i) наличие ПШ на словоразделе наиболее явно зависит от присутствия границы финитной клаузы и маркера бессоюзного сочинения; ii) отдельные синтаксические конструкции связаны с ПШ определенного уровня глубины; iii) при учете скобочной разметки непосредственных составляющих число закрывающих скобок на словоразделе является признаком, в наибольшей степени коррелирующим с наличием ПШ.

4. Заключение

Целью данной работы было рассмотрение синтаксических факторов, влияющих на наличие просодической границы. Авторы ставили перед собой задачу прежде всего обнаружить такие факторы, детальное обсуждение их природы и причин влияния синтаксиса на просодию мы оставляем на будущее.

Как мы видели, в целом ряде случаев наблюдается устойчивая связь между синтаксисом и просодией. Отмеченные факты зависимости просодической структуры от синтаксической, очевидно, должны учитываться при создании систем синтеза речи и решении других похожих задач.

Отметим, что для получения более надежных свидетельств о характере связи синтаксиса и просодии в последующих экспериментах необходимо расширить выборку для отдельных синтаксических конструкций, ввести дополнительную разметку и т. д. Для более чистого эксперимента, исключающего влияние пунктуации и индивидуальных особенностей диктора, необходим также анализ синтаксиса и просодии (в корпусах) спонтанной речи.

Литература

1. *Cinque G.* (1993), A null theory of phrase and compound stress, Linguistic Inquiry, 24, pp. 239–298.
2. *Chistikov P. G., Khomitsevich O. G., Rybin S. V.* (2014), Statistical methods for automatic determination of pause locations and length in TTS systems [Statisticheskiye metody avtomaticheskogo opredeleniya mest i dlitel'nosti pauz v sistemakh sinteza rechi], Proceedings of Higher Schools. Instrument engineering [Izvestiya vuzov. Priborostroyeniye], 57(2), pp. 28–32.
3. *Croft W.* (1995), Intonational Units and Grammatical Units, Linguistics, 33, pp. 839–882.
4. *Dyachenko P. V., Iomdin L. L., Lazursky A. V., Mityushin L. G., Podlesskaya O. Yu., Sizov V. G., Frolova T. I., Tsinman L. L.* (2015), A Deeply Annotated Corpus

Of Russian Texts (SynTagRus): Contemporary State Of Affairs [Sovremennoe Sostojanie Gluboko Annotirovannogo Korpusa Tekstov Russkogo Jazyka (SinTagRus)], Proceedings of the V. V. Vinogradov Russian Language Institute [Trudy Instituta Russkogo Yazyka imeni V. V. Vinogradova] Moscow, 2015, v. 6. pp. 272–299.

5. *Fach M. L.* (1999), A comparison between syntactic and prosodic phrasing, Proceedings of EUROSPEECH'99, pp. 527–530.
6. *Grashchenkov P, Kirillova A, Smirnova O. S.* (2018) Syntactic Factors That Influence Prosody [Sintaksicheskie faktory, vlijajuschie na prosodiju], Kompjuternaja Lingvistika i Intellektual'nye Texhnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2018" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018"], http://www.dialog-21.ru/...
7. *Halliday M. A. K.* (1967), Notes on transitivity and theme in English. Part 2, Journal of Linguistics, 3, pp. 199–244.
8. *Iomdin L. L., Lobanov B. M.* (2009). Syntactic correlates of prosodically marked elements of the sentence and their role in the tasks of-text-to-speech synthesis [Sintaksicheskie korreljaty prosodicheski markirovannyh elementov predlozhenija i ih rol' v zadachah sinteza rechi po tekstu], Kompjuternaja Lingvistika i Intellektual'nye Texhnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2009" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2009"], Moscow, pp.: 136–142.
9. *Iomdin L. L., Lobanov B. M., Getsevich Iu. S.* (2011). The talking ETAP. Using the ETAP parser in Russian speech synthesis [Govorjashhij «ETAP». Opyt ispol'zovanija sintaksicheskogo analizatora sistemy ETAP v russkom rechevom sinteze] Kompjuternaja Lingvistika i Intellektual'nye Texhnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2011" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2011"], Moscow, pp.: 269–279.
10. *Kodzasov S. V., Krivnova O. F.* (2001), General phonetics [Obschaya fonetika], Russian State University for the Humanities, Moscow.
11. *Kodzasov S. V.* (2009), Research in the field of Russian prosody [Issledovaniya v oblasti russkoy prosodii], Yazyki slavyanskikh kul'tur, Moscow.
12. *Krivnova O. F., Chardin I. S.* (1999), Pausation in automatic speech synthesis [Pauzirovaniye pri avtomaticheskom sinteze rechi], Theory and practice in speech research. Proceedings of ARSO-99 [Teoriya i praktika rechevykh issledovaniy (ARSO-99). Materialy konferenzii], Moscow, pp. 87–103.
13. *Krivnova O. F.* (2015), The Depth of Prosodic Breaks in Spoken Text (Experimental Data) [Glubina prosodicheskikh shvov v zvuchaschem tekste (eksperimental'niye danniye)], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2015" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2015"], Moscow, pp. 338–351.
14. *Khomitsevich O. G., Solomennik M. V.* (2010), Automatic pause placement in a Russian text-to-speech system [Avtomaticheskaya rasstanovka pauz v sisteme sinteza russkoy rechi po tekstu], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2010"

[Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2010"], Bekasovo, pp. 531–537.
15. *Khomitsevich O. G., Chistikov P. G.* (2013), Using statistical methods for prosodic boundary detection and break duration prediction in a Russian TTS system, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2013" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2013"], Bekasovo, pp. 2–11.
16. *Lahiri A., Plank F.* (2010), Phonological phrasing in Germanic: The judgement of history, confirmed through experiment, Transactions of the Philological Society, 108(3), pp. 370–398.
17. *Lobanov B. M.* (2008), An algorithm of text segmentation on syntactic syntagmas for TTS synthesis [Algoritm segmentazii teksta na sintaksicheskiye sintagmy dlya sinteza rechi], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2008" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2008"], Moscow, pp. 323–329.
18. *Paducheva E. V.* (2016), Communicative structure and Linear-Accent Transformations [Kommunikativnaya struktura i lineyno-akzentniye preobrazovaniya (na materiale russkogo yazyka)], in A. V. Zimmerling, E. A. Lyutikova (eds.), Clause Architecture in the Parametric Models: Syntax, Information Structure, Word Order [Arkhitektura klauzy v parametricheskikh modelyakh: sintaksis, informazionnaya struktura, poryadok slov], Yazyki slavyanskikh kul'tur, Moscow, pp. 25–75.
19. *Sanderman A. A.* (1996), Prosodic phrasing: Production, perception, acceptability, and comprehension, Unpublished doctoral dissertation, University of Eindhoven, The Netherlands.
20. *Selkirk E.* (1978), On prosodic structure and its relation to syntactic structure, in T. Fretheim (ed.), Nordic Prosody II: Papers from a symposium, Trondheim: TAPIR, pp. 111–140.
21. *Smirnova O. S.* (2017), Statistical Analysis of Perceptive Estimation for Depth of Prosodic Breaks in Russian Spoken Text [Statisticheskiy analiz rezul'tatov perzeptivnogo ozenivaniya glubiny prosodicheskikh shvov v russkom zvuchaschem tekste], paper presented at the International Conference "Dialog 2017", Moscow.
22. *Steedman M.* (2000), Information Structure and the Syntax-Phonology Interface, Linguistic Inquiry, 34, pp. 649–689.
23. *Tepperman J., Nava E.* (2011), Where Should Pitch Accents and Phrase Breaks Go? A Syntax Tree Transducer Solution, Proceedings of INTERSPEECH-2011, pp. 1353–1356.
24. *Yanko T. E.* (2008), Intonational strategies of Russian speech in a comparative aspect [Intonazionniye strategii russkoy rechi v sopostavitel'nom aspekte], Yazyki slavyanskoy kul'tury, Moscow.

НАДКОРПУСНАЯ БАЗА ДАННЫХ КАК ИНСТРУМЕНТ ИЗУЧЕНИЯ ФОРМАЛЬНОЙ ВАРИАТИВНОСТИ КОННЕКТОРОВ[1]

Инькова О. Ю. (Olga.Inkova@unige.ch)

ИПИ ФИЦ ИУ РАН, Москва, Россия;
Женевский университет, Женева, Швейцария

Ключевые слова: базы данных коннекторов, русский язык, корпусная лингвистика, структура коннекторов, формальная вариативность, статистические данные

SUPRACORPORA DATABASE AS AN INSTRUMENT OF THE STUDY OF THE FORMAL VARIABILITY OF CONNECTIVES

Inkova O. Yu. (Olga.Inkova@unige.ch)

Institute of Informatics Problems, FRC CSC RAS, Moscow, Russia;
University of Geneva, Geneva, Switzerland

The article intends to describe the formal variation of the connectors of the Russian language on the basis of a cognitive-semantic approach. Every *discourse variant DV* of a connector K, *i.* e. the specific form assumed by K in a discourse section, is singled out, and registered in the supracorpora database of connectors (*SCDB*), in which a system of intersecting clusters has been developed, allowing to assign in the course of the annotation the same *DV* to different structural clusters. In the next phase, on the base of further semantic analysis, the *DV*s with a common element are combined into a structural-semantic complex around a *basic form*: the minimal linguistic unit that enables the speaker to express a certain logical-semantic relation, and the listener to identify it. In conclusion, criteria for describing the formal variation of the connectors are proposed, as well as examples of the "profiles" of the basic forms. They reflect the potential of linguistic means that the speaker has at his disposal to express one or another logical-semantic relations or one of their combinations.

Key words: supracorpora database of connectives, Russian, corpus linguistics, formal variability of connectives, structure of connectives, statistical data

[1] Исследование выполнено в ИПИ ФИЦ ИУ РАН при поддержке РФФИ (грант № 16-06-00070).

1. Введение

Хорошо известно, что состав связующих средств русского языка продолжает пополняться, как за счет новых образований из «блуждающих частиц» (Николаева 1985, 2008), так и за счет возникновения новых сочетаний из уже существующих языковых единиц, в частности союзов с частицами, модальными словами, наречиями и др. В связи с этим возникает вопрос: правомерно ли рассматривать каждое из них как отдельную языковую единицу? Представляется очевидным, что простой союз иначе соотносится с возникшими на его основе сочетаниями, чем с другим простым союзом. Иначе говоря, место в системе русского языка, с одной стороны, простых союзов *и — да — а — но — или*, а с другой, образований типа *да — да и — да еще и, да притом еще, да притом еще и, да еще притом, да вдобавок еще* и т. д., по-видимому, различно (см. постановку проблемы в Прияткина 1977, Черемисина, Колосова 2009: 123–139).

Словари служебных слов и «сочетаний, эквивалентных слову» исходят из упрощенной трактовки проблемы, фиксируя в отдельной словарной статье — причем на довольно субъективной основе — некоторые из этих комбинаций и оставляя за бортом другие, имеющие, на наш взгляд, такой же статус. Так, из перечисленных выше сочетаний с *да* фиксируются, как правило (см. сводные таблицы в Богданов, Рыжова 1997), *да, да и, да еще* и *да еще и*. При этом [Морковкин 1997], [Бурцева 2010], [Рогожникова 2003] рассматривают *да еще и* как вариант *да еще*, а [Ефремова 2004] посвящает каждому из них словарную статью. [Рогожникова 2003] фиксирует также форму *да вдобавок еще*. Еще сложнее обстоит дело с двухместными коннекторами: например, если для *хотя* ни один из четырех словарей не фиксирует возможность образовывать двухместный коннектор, то для *едва* [Ефремова 2004] и [Рогожникова 2003] дают форму *едва… как*, а [Морковкин 1997] — *едва… а*.

Схожая позиция характеризует корпуса русского языка. Хельсинский аннотированный корпус русских текстов (ХАНКО) опирается в этом вопросе на список служебных многокомпонентных единиц, которые считаются «фраземами» [Мустайоки, Копотев 2004]. Он создан разработчиками корпуса на основе словарей и используется для морфологической разметки. В списке союзов 323 единицы.

В НКРЯ создан «Корпусный словарь неоднословных лексических единиц», которые квалифицируются как «устойчивые лексические обороты» (http://ruscorpora.ru/obgrams.html). Необходимость создания такого словаря объясняется тем, что морфологическая разметка дается в НКРЯ для орфографического слова, т. е. отделяемого пробелом. В основу словаря положены данные частотных коллокаций НКРЯ с дополнениями из словаря [Рогожникова 2003] и МАС. Список оборотов в функции союза и союзного слова включает 159 единиц (т. е. меньше половины списка ХАНКО), для каждой из которой дано ориентировочное количество употреблений. В число «устойчивых» попадают обороты *добро бы еще* и *диви б* с единичными употреблениями; фиксируются как самостоятельные лексические единицы *добро бы* и *добро б*, *если бы* и *если б*, тогда как из перечисленных выше комбинаций с *да* фиксируются только *да и* и *да еще*, хотя поиск в НКРЯ по *для притом еще* выдает 57 вхождений, по *да вдобавок*

еще — 50. Двухместных коннекторов в списке нет, хотя некоторые логико-семантические отношения (ЛСО) выражаются только ими (аналогия: *как… так*, сопоставительное *если… то* и др.). С другой стороны, в списке союзов и союзных слов оказываются такие единицы, как *что касается* и *что до*: их оценка как показателей связи требует теоретического обоснования.

Оба корпуса (НКРЯ и ХАНКО) считают многокомпонентные или неоднословные единицы устойчивыми сочетаниями, фразеологизмами, эквивалентами слова, иначе говоря, единицами, принадлежащими языку, отсюда акцент на их устойчивость и частотность. Такой подход приводит к определенной бессистемности и субъективности в описании состава служебных единиц русского языка, поскольку критерий частотности применяется непоследовательно, а критерий устойчивости не определен с достаточной четкостью[2].

2. Когнитивно-семантический подход к проблеме формального варьирования коннекторов

Для описания структуры коннекторов, состоящих более чем из одного ‘слова’, мы предлагаем использовать когнитивно-семантический подход, позволяющий дать адекватное отражение способности носителя языка выражать или устанавливать при восприятии речи то или иное ЛСО между ситуациями. Это означает, прежде всего, разграничение понятийного и языкового уровня анализа (соответственно, тип ЛСО и языковые средства, которыми он может выражаться), поскольку выбор коннектора и его состава для выражения ЛСО зависит не только от знаний языка, но и от коммуникативного намерения говорящего, в передаче которого задействованы также общие принципы общения и знания о мире. Чем типичней сценарий, лежащий в основе описываемых событий, тем меньше необходимость выражать между ними отношения соответствующим показателем (ср. *Наша Таня громко плачет: уронила в речку мячик*); с другой стороны, чем важнее с коммуникативной точки зрения данное ЛСО для говорящего, тем тяжеловесней будет его показатель: ср. ниже три примера с *как только*. Если говорящий решает эксплицировать ЛСО, то дальше в силу вступают семантические законы, определяющие возможности сочетаний языковых единиц и состав показателя ЛСО.

При таком подходе ‘неоднословные’ коннекторы, обладающие свойством вариативности формы, рассматриваются не как целостные грамматические единицы, входящие в знание языка и воспроизводимые в готовой форме, а как свободные сочетания, принадлежащие сфере речи. Они, в частности, обладают важнейшим для автономности составляющих их словоформ свойством

[2] В связи с созданием корпусов, посвященных дискурсивным, или риторическим, в терминологии данного направления исследований, отношениям (ср., например, *The Penn Discourse Treebank*), возрос интерес и к их показателям, в частности, к тем случаям, когда в одном высказывании присутствует не один, а несколько маркеров риторических отношений (см., например, Grote & al. 1995, Webber & Joshi 1998, Oates 2000, Fraser 2013, Webber 2016). Однако внимание в этих работах уделяется преимущественно самим отношениям, и вопрос о формальной вариативности коннекторов не ставится.

отделимости, допуская вставку между ними других слов [Инькова 2016, Кобозева 2016]. Ср.: *Нет я усядусь на скамеечке.* ***Да и*** *что я стану там делать* и *Звонить зря было нечего,* ***да*** *ему* ***и*** *не к фигуре.*

Основной единицей анализа при когнитивно-семантическом подходе становится *речевая реализация* (РР), т. е. та форма, в которой коннектор встречается в данном конкретном высказывании. Именно она регистрируется в надкорпусной базе данных коннекторов (НБД). Ср. три РР для *как только*, выражающие ЛСО контактного предшествования: ***Как только*** *дело дошло до «честного слова», я махаю руками и сажусь за стол* (Чехов); ***Как только*** *кончается отношение служебное,* ***так*** *кончается всякое другое* (Л. Толстой); *Он,* ***как только*** *проснулся,* ***тотчас же*** *вознамерился встать, умыться* (Гончаров).

3. НБД как инструмент анализа структуры коннекторов

Применяемая в НБД система аннотирования (подробнее см. Inkova, Popkova 2017) позволяет решать теоретически сложный вопрос о том, являются ли приведенные выше три РР с общим элементом *как только* вариантами одного коннектора (и какого именно) или разными коннекторами не на этапе аннотирования, а после семантического анализа соответствующей группы РР. На первом этапе аннотирования каждая регистрируемая РР приписывается к двум типам кластеров. В четырех кластерах первого типа учитывается количественный состав РР, а также наличие показателя ЛСО в одном или в каждом из соединяемых фрагментов текста:

- одноэлементные: состоящие из одного элемента, т. е. РР, неразложимые на более мелкие языковые единицы, например, *и*, *или*, *но*, *а*, *хотя*, *тогда*, *иначе*, *словом*;
- многоэлементные: РР, состоящие из нескольких элементов, например, *и вообще*, *к тому же*, *но при всем при том*, *еще и*|*вдобавок*[3];
- двухкомпонентные: РР, компоненты которых вводят два текстовых фрагмента, например, *даже если бы*||*то*[4], *как только*||*тут же*, *хотя*|*и*||*но*|*между тем*;
- многокомпонентные: РР, компоненты которых вводят более двух текстовых фрагментов (например, *не только*||*но даже и*||*и даже, хотя*||*хотя*||*однако все же*).

Заметим, что из 926 РР, зарегистрированных в НБД на 01.02.2018, на долю одноэлементных приходится чуть более 5% (47), многоэлементных — почти 50% (457), двухкомпонентных — около 40% (357) и многокомпонентных — чуть более 5% (47).

---

[3] Одиночная вертикальная черта в названии РР означает, что элементы РР разделены другими словами.

[4] Двойная вертикальная черта в названии двух- и многокомпонентных РР коннекторов означает, что компоненты не только разделены другими словами, но и вводят разные текстовые фрагменты.

Перечисленные кластеры первого типа, отражая важный классификационный признак коннекторов (их состав), не учитывают тот факт, что одна и та же языковая единица может входить в различные сочетания, см. выше примеры для *как только*. Поэтому описание состава многоэлементных, двух- и многокомпонентных РР и сочетаемости составляющих их компонентов или элементов решается в НБД через систему «перекрестных», или «пересекающихся», кластеров второго типа, которые позволяют разметчикам одновременно помещать аннотируемую РР в разные кластеры и которые пополняются в процессе аннотирования. См. Таблицы 1 и 2.

Таблица 1. Примеры РР с указанием кластеров первого и второго типов, в которые они входят (состояние на 1 февраля 2018 г.)

	Речевые реализации (РР)		
	а впрочем	***да еще вдобавок***	***не только || а даже и***
Кластеры РР	«Многоэлементные»	«Многоэлементные»	«Двухкомпонентные»
	«а»	«вдобавок»	«даже»
	«впрочем»	«да»	«а»
		«еще»	«и»
			«не»
			«не только»
			«только»

Таблица 2. Примеры кластеров второго типа и входящих в них РР (состояние на 1 февраля 2018 г.)

	Кластеры РР		
	«вдобавок»	**«значит»**	**«притом»**
РР, входящие в указанные кластеры	вдобавок	а значит	да и притом
	да вдобавок	если||значит	да притом
	да еще вдобавок	если только||значит	да притом еще
	еще вдобавок	значит	и|притом
	еще и|вдобавок	коли||то значит	и притом
	и вдобавок		

На основе последующего семантического анализа зарегистрированные в НБД РР объединяются в структурно-семантический комплекс вокруг *базовой формы* (БФ), т. е. той минимальной РР, которая позволяет говорящему выразить некоторое ЛСО, а слушающему — его идентифицировать. Например, для выражения отношения контактного предшествования говорящий может использовать РР *лишь*, а также целый ряд РР, включающих этот элемент: *лишь*||*как*; *едва лишь*; *едва лишь*||*как*; *только лишь*; *только лишь*||*как*; *лишь только*; *лишь только*||*как*; *чуть лишь*; *чуть лишь*||*как*; *едва лишь только*; *едва лишь*

только||как и ряд других [Кобозева 2017]. Форма *лишь* будет считаться БФ, поскольку она может выражать это ЛСО сама по себе. Аналогичным образом, выделяются кластеры, объединенные вокруг БФ *едва*, *только* и *чуть*, пересекающиеся друг с другом благодаря наличию РР, включающих две БФ.

БФ удобно использовать и для номинации кластера РР, который можно рассматривать как абстрактную единицу системы коннекторов данного языка. Поэтому в НБД наряду, например, с кластером *только* создан кластер *не только*: последний является также именем БФ, выражающей отношение неединственности [Инькова 2016].

Статистический анализ употреблений РР позволил заметить, что БФ обычно далеко не самая частотная РР в своем кластере. Наиболее частотную по корпусным данным РР для выражения того или иного ЛСО предлагается называть *основной формой коннектора*. Так, в кластере *лишь* основной формой будет РР *лишь только*. В кластере *не только* — РР *не только||но и* [Амеличева 2017].

Описание формальной вариативности неоднословных коннекторов предлагается делать исходя из описанных выше теоретических принципов и с учетом приводимых ниже параметров, характеризующих БФ коннектора.

1. Возможность для БФ образовывать многоэлементные РР, выражающие то же ЛСО.

Доп. критерий:

- контактное vs. дистантное расположение элементов.

2. Возможность для БФ образовывать двухкомпонентные РР.

Доп. критерии:

- симметричность vs. асимметричность компонентов,
- вариативность компонентов (возможность образовывать многоэлементные компоненты),
- контактное vs. дистантное расположение элементов.

3. Возможность для БФ образовывать многокомпонентные РР.

Доп. критерии:

- симметричность vs. асимметричность компонентов,
- вариативность компонентов (возможность образовывать многоэлементные компоненты).

4. Возможность для БФ сочетаться с показателями других ЛСО.

Доп. критерии:

- ЛСО той же семантической группы,
- ЛСО другой семантической группы.

Классы РР, полученные на основе этих параметров благодаря информации из НБД, будут «перекрестными», но они позволяют очертить семантический и формальный профиль БФ показателя того или иного ЛСО. Несколько таких профилей приведены ниже в Таблице 3. Важно подчеркнуть, что речь идет не о лексических единицах, существующих в языке a priori, а именно о потенциале языковых средств, которым располагает говорящий для выражения

ЛСО[5], что отражено в использовании слова *возможность* при формулировании параметров.

Например

1. Данная БФ для выражения ЛСО спецификации не может изменять свой состав для выражения того же отношения. В этом она схожа с БФ *зато* и отличается от БФ *также*, характеризующейся формальной вариативностью.

2–3. *Например* не может образовывать ни двух-, ни многокомпонентные РР для выражения того же ЛСО.

4. Данная БФ сочетается с показателями других ЛСО, относящихся как к той же семантической группе, так и к другим. Ср. (1) в сочетании с другим показателем спецификации *в частности*, и (14) ниже с показателем ЛСО сопоставления:

(1) Под флагами повсеместно висели листовки с «обещаниями» фюрера своим рабочим: два десятка пунктов, *в частности, например*, обещание сделать 1 мая Общенациональным праздником немецких трудящихся и оплаченным выходным днем. [Елена Съянова. Лей — обольститель немецкого рабочего класса // «Знание — сила», 2005]

Также

1. Данная БФ может образовывать многоэлементные РР для выражения того же ЛСО (аддитивность), как с контактным (3), так и с дистантным (4) расположением элементов.

---

[5] Для названия ЛСО в данной работе намеренно используется нейтральная терминология русской грамматической традиции, дополняемая при отсутствии необходимого термина, рубриками из классификации, разрабатываемой автором статьи. Этот терминологический компромисс обусловлен тем, что существующие на сегодняшний день классификации, используемые для разметки корпусов и в основе которых лежит, как правило, Теория риторической структуры (RST), не могут быть использованы для анализа внутренней формы коннекторов и их сочетаемости. С одной стороны, в них не хватает рубрик, чтобы различить, например, ЛСО, выражаемые *а*, *зато* и *наоборот*; ср. предложение «мирового стандарта» отношений в (Prasad & Bunt 2016), включающий 20 отношений: *Cause, Condition, Negative Condition, Purpose, Manner, Concession, Contrast, Exception, Similarity, Substitution, Conjunction, Disjunction, Exemplification, Elaboration, Restatement, Synchrony, Asynchrony, Expansion, Functional dependence, Feedback dependence*, или, наоборот, чрезвычайно детальную классификацию, включающую 62 риторических отношения, но тем не менее слабо разработанную в зоне ЛСО, в [Кибрик, Подлесская 2009]. С другой стороны, все они содержат «риторические отношения», многие из которых не являются логико-семантическими; ср. в «мировом стандарте», например, последние два или в классификации А. Кибрика и В. Подлесской *Begin, Emotional reaction, Headline, Problem* и некоторые другие. Разрабатываемая автором классификация предназначена именно для анализа ЛСО, выражаемых коннекторами. Она основана на семантических операциях, лежащих в основе отношения, и последовательно различает уровни, на которых ЛСО может быть установлено (пропозиция, высказывание, метатекст); см. обзор существующих классификаций и основные теоретические положения, лежащие в основе предлагаемой классификации, в [Inkova 2017].

Таблица 3. Примеры профилей БФ коннекторов

Основной параметр	**Дополнительный параметр**	*например*	*также*	*зато*	*если… то* **(сопоставительный)**	*не только*	*или*
1. Возможность образовывать однокомпонентные многоэлементные РР, выражающие то же ЛСО	контактное расположение	–	+	–	–	–	+
	дистантное расположение	–	+	–	–	–	–
2. Возможность образовывать двухкомпонентные РР, выражающие то же ЛСО	симметричность vs. асимметричность компонентов	–	–	+ симметричность	+ асимметричность	+ асимметричность	+ симметричность + асимметричность
	вариативность компонентов	–	–	–	–	–	+ второй компонент
	контактное vs. дистантное расположение элементов	–	–	–	–	–	+ второй компонент
3. Возможность образовывать многокомпонентные РР, выражающие то же ЛСО	симметричность vs. асимметричность компонентов	–	–	+ симметричность	+ асимметричность	+ асимметричность	+ симметричность + асимметричность
	вариативность компонентов	–	–	–	–	–	+ второй компонент
4. Возможность сочетаться с показателями других ЛСО	ЛСО той же семантической группы	+	+	+	–	–	+
	ЛСО другой семантической группы	+	+	+	+	+	+

(2) (…) около него красовались лубочные картинки, представляющие взятие Кистрина и Очакова, *также* выбор невесты и погребение кота [А. С. Пушкин. Капитанская дочка (1836)]

(3) Выяснилось, что в бюро иностранцев ни о каком Воланде, *а равно также и* Фаланде, маге, ровно ничего не слыхали. [М. А. Булгаков. Мастер и Маргарита, (1929–1940)]

(4) Мы не будем здесь останавливаться на деталях обряда увенчания — развенчания (хотя они очень интересны) и на различных вариациях его по эпохам и разным празднествам карнавального типа. Не будем *также* анализировать *и* различные побочные обряды карнавала. [М. М. Бахтин. Проблемы поэтики Достоевского (1963)]

2–3. Данная БФ не может образовывать двух- и многокомпонентные РР.

4. *Также* может сочетаться с показателями того же ЛСО, например, с градационными показателями аддитивных пропозициональных ЛСО, как *сверх того* в (5), и с показателями ЛСО других семантических групп; ср. в (6) с показателем отношения неединственности *не только||но*:

(5) Тургенев мне все более и более нравится: у него ум — необыкновенный ум, *сверх того также* чутье, способность к анализу, уменье хорошо подступиться ко всякой вещи. [А. В. Дружинин. Дневник (1845)]

(6) Самоубийство есть *не только* ложное и греховное отношение к жизни, *но также* ложное и греховное отношение к смерти. [Н. А. Бердяев. О самоубийстве (1931)]

Зато

1. Как и БФ *например*, БФ *зато* не может изменять свой состав для выражения того же ЛСО (возместительное противопоставление).

2–3. В отличие от *например* и *также*, БФ *зато* для выражения того же ЛСО может образовывать двух- (7) и многокомпонентные (8) РР, но только с симметричными компонентами (повторяющееся два раза или более *зато*):

(7) У него не было идолов, *зато* он сохранил силу души, крепость тела, *зато* он был целомудренно-горд. [И. А. Гончаров. Обломов (1859)]

(8) Ни жеманства, ни кокетства, никакой лжи, никакой мишуры, ни умысла! *Зато* ее и ценил почти один Штольц, *зато* не одну мазурку просидела она одна, не скрывая скуки; *зато*, глядя на нее, самые любезные из молодых людей были неразговорчивы, не зная, что и как сказать ей… [И. А. Гончаров. Обломов (1859)]

4. БФ *зато* может сочетаться с показателями других ЛСО, той же семантической группы и других групп. В (9) *зато* образует многоэлементную РР с дистантным расположением элементов с показателем ЛСО той же семантической группы *наоборот* (контрастное противопоставление):

(9) В последние два года средства от дорожных штрафов несколько сократились, *зато* аппетиты чиновников, *наоборот*, возросли. [Сергей Ждакаев. Ильич для камуфляжа. Кто кормился из внебюджетного фонда (2001) // «Известия», 2001.07.18]

В (10) *зато* сочетается с показателем уступительного ЛСО *хоть и* в составе двухкомпонентной РР:

(10) Дай же, я думаю, *хоть и* упущу на время одно, *зато* другое схвачу за хвост, — своего-то, своего-то, по крайности, не упущу. [Ф. М. Достоевский. Преступление и наказание (1866)]

***Если… то* (сопоставительный)**

1–2 Одиночный союз *если* не может выражать ЛСО сопоставления, поэтому БФ для этого ЛСО будет считаться двухкомпонентная *если… то.* Ср. (11) и (12)

(11) *Если* в Александровском округе климат морской, *то* в Тымовском он континентальный [А. П. Чехов. Остров Сахалин. Пример (БАС)]

(12) ??*Если* в Александровском округе климат морской, в Тымовском он континентальный

Если… то сопоставления не может менять свой состав для выражения того же ЛСО.

3. Можно предположить, хотя такие примеры пока не зафиксированы в НБД, что данная БФ для выражения того же ЛСО может входить в состав многокомпонентных РР с повторяющимся первым компонентом *если* и вторым компонентом *то*:

(13) *Если* черно-белый котенок всего боялся и жалобно мяукал, *если* рыжий все время от нас прятался, *то* полосатый сразу почувствовал себя как дома.

4. БФ *если… то* не может сочетаться с другими показателями сопоставительных отношений, но может сочетаться с показателями других ЛСО, которые будут входить, как правило, в состав второго компонента: ср. (14) с показателем ЛСО спецификации:

(14) *Если* в Москве Лужков решительно отказал в этом МГТС, *то, например*, в Нижневартовске «повременка» будет внедрена до конца 2003 г. [Михаил Классон. Тарифные аппетиты (2003) // «Время МН», 2003.05.26]

Не только

1. БФ *не только* выражает ЛСО неединственности:

(15) Он до того углубился в себя и уединился от всех, что боялся даже всякой встречи, *не только* встречи с хозяйкой. [Ф. М. Достоевский. Преступление и наказание (1866)]

Она может расширять свой минимальный состав, образуя многоэлементные РР:

(16) Важно раскрыть функцию идей в полифоническом мире Достоевского, *а не только* их монологическую субстанцию. [М. М. Бахтин. Проблемы поэтики Достоевского (1963)]

2. БФ *не только* может выражать то же ЛСО в составе двухкомпонентных РР, но, в отличие от *если… то*, характеризуется формальным варьированием во втором компоненте. Ср. (17), и (18):

(17) В деревне, без службы Иван Ильич в первый раз почувствовал *не только* скуку, *но* тоску невыносимую. [Л. Н. Толстой. Смерть Ивана Ильича (1886)]

(18) *Не только* драньем вихров, *но даже и* помелом было бы полезно обойтись с иными дураками. [Ф. М. Достоевский. Преступление и наказание (1866)]

3. БФ *не только* может входить в состав многокомпонентных асимметричных РР с формальным варьированием в компонентах, не включающих БФ.

(19) И *не только* потому, что размещался он в двух больших залах со сводчатыми потолками, расписанными лиловыми лошадьми с ассирийскими гривами, *не только* потому, что на каждом столике помещалась лампа, накрытая шалью, *не только* потому, что туда не мог проникнуть первый попавшийся человек с улицы, *а еще и* потому, что качеством своей провизии Грибоедов бил любой ресторан в Москве, как хотел, *и* что эту провизию отпускали по самой сходной, отнюдь не обременительной цене. [М. А. Булгаков. Мастер и Маргарита (1929–1940)]

4. БФ *не только* может сочетаться с показателями ЛСО, относящихся к другим семантическим группам: в (20) — с показателем ЛСО контрастного противопоставления *напротив*:

(20) <…> и это объясняется *не только* его положением журналиста, требующим трактовки всего в разрезе современности; *напротив*, мы думаем, что пристрастие Достоевского к журналистике и его любовь к газете <…> объясняются именно основною особенностью его художественного видения. [М. М. Бахтин. Проблемы поэтики Достоевского (1963)]

Или

1. БФ *или* может выражать ЛСО альтернативы в своем минимальном составе, допуская незначительное формальное варьирование в сочетании с *же* (21):

(21) Как будто дело происходит в гастрономе *или же* на рынке. [Сергей Довлатов. Иностранка (1985)]

2–3. Она может выражать то же ЛСО в составе двух- и многокомпонентных РР, как симметричных (повторяющееся *или*), так и асимметричных (22). В составе таких РР БФ *или* может варьироваться (как правило, заключительный компонент ряда).

(22) Требовалось выяснить, были *ли* похищены эти женщины шайкой убийц и поджигателей *или же* бежали вместе с преступной компанией добровольно? [М. А. Булгаков. Мастер и Маргарита (1929–1940)]

4. БФ *или* может сочетаться с показателями других ЛСО, относящихся как к той же семантической группе, например, разделительное *а то* в (23), так и к другим семантическим группам; ср. (24) с показателем ЛСО коррекции *вернее*:

(23) Возьмите две капельки амбре, одну капельку вервены и получите дух настоящий… настоящий, — она пожевала губами, ища слова, — земной и небесный. *А то* возьмите основной дух Трефль инкарнат, пряный, точно с корицей, да в него на три капли одну белого ириса… <…> *Или* возьмите нежную Икзору, — не слушая, продолжает фантазировать мадам Лазенская, — а к ней подлейте одну каплю тяжелого Фужеру… [Н. А. Тэффи. За стеной (1910)]

(24) Основное, что определяло его лицо, это было, пожалуй, выражение добродушия, которое нарушали, впрочем, глаза, *или, вернее*, не глаза, а манера пришедшего глядеть на собеседника. [М. А. Булгаков. Мастер и Маргарита (1929–1940)]

4. Выводы

Предлагаемый когнитивно-семантический подход к проблеме формального варьирования коннекторов дает, на наш взгляд, более адекватное представление о механизмах выражения того или иного ЛСО и о том потенциале языковых средств, которым располагает для этого говорящий. Регистрируемые в НБД РР являются при таком подходе единицами речи, каждый раз ex novo создаваемыми говорящим с учетом его коммуникативного намерения и на основе семантических законов, определяющих возможности сочетаемости входящих в их состав элементов. Разработанная в НБД система аннотирования с использованием перекрестных кластеров позволяет решать теоретически сложный вопрос о статусе регистрируемой РР не в момент ее аннотирования, а на основе ее дальнейшего семантического анализа. Особый интерес представляет соотношение состава РР русского языка и частотности их употребления. Если на долю одноэлементных РР приходится, как мы видели всего 5% (47) от общего числа зарегистрированных РР, то доля их употребления в текстах составляет 38% (они были зарегистрированы в 2684 аннотациях из 7016). Наиболее частотными по данным НБД являются многоэлементные РР (42.5% или 2983 аннотации), наименее частотными — многокомпонентные (0.8%). Употребление двухкомпонентных РР составляет 16.9%. НБД является, таким образом,

эффективным инструментом, позволяющим судить как о формальном составе связующих средств русского языка, так и о частоте их употребления в тексте, и исчислить их возможные варианты.

Литература

1. *Amelicheva V.* (2017). Formal'noe i semanticheskoe var'irovanie russkogo konnektora *ne tol'ko… no i* i ego frantsuzskie ekvivalenty [Formal and semantic variations of the Russian connector *ne tol'ko… no i* and its French equivalents]. Contrastive linguistics. 2017. Vol. XLII, No. 4. Pp. 9–20.
2. *Bogdanov S. I., Ryzhova Yu. V.* (1997). Russkaya sluzhebnaya leksika [Russian structural words], St. Petersburg, St. Petersburg State Univ.
3. *Burtseva V. V. ed.* (2010). Slovar' narechii i sluzhebnykh slov russkogo yazyka [Dictionary of the adverbs and structural words of Russian language], Moskva, Drofa.
4. *Cheremisina M. I., Kolosova T. A.* (2010). Ocherki po teorii slozhnogo predlozheniya [An outline of the complex sentence theory], Moscow, URSS. (1st ed. Novosibirsk, Nauka, 1987).
5. *Efremova T. F.* (2004). Tolkovyi slovar' sluzhebnykh chastei rechi russkogo yazyka [Explanatory dictionary of the Russian auxiliary parts of speech], Moscow, Astrel'-Ast.
6. *Fraser B.* (2013). Combinations of Contrastive Discourse Markers in English. International Review of Pragmatics 5. 2013. Pp. 318–340.
7. *Grote B., Lenke N., Stede M.* (1995). Mar(k)ing Concessions in English and German. Proceedings of the 5th European Workshop on Natural Language Generation, Leiden, May, Leiden University. Pp. 11–32.
8. *HANCO* — The Helsinki Annotated Corpus, available at: http://www.ling.helsinki.fi/projects/hanco/index_e.html.
9. *Inkova O.* (2016). K probleme opisanija mnogokomponentnykh konnektorov russkogo yazyka [On the problem of the description of multiword connectives of Russian language: *ne tol'ko… no i* (no only… but also)]. Voprosy jazykoznanija. 2016. No. 2. Pp. 37–60.
10. *Inkova O.* (2017). Le relazioni logico-semantiche tra gli enunciati: una proposta di classificazione [The logical-semantic relations: a classification proposal]. M. di Filippo, F. Esvan (eds). Studi di linguistica slava. Napoli, Il Torcoliere. Pp. 105–124.
11. *Inkova O., Popkova N.* (2017). Statistical data as information source for linguistic analysis of Russian connectors. Informatics and applications. 2017. Vol. 11, No. 3. Pp. 123–131.
12. *Kibrik A. A., Podlesskaya V. I.* (2009). Rasskazy o snovideniyakh. Korpusnoe issledovanie ustnogo russkogo diskursa [Night dream stories: A corpus study of Russian spoken discourse]. Moscow, Yazyki Slavyanskikh Kul'tur.
13. *Kobozeva I. M.* (2016). Kognitivno-semanticheskii podkhod k opisaniju sredstv svjazi predlozhenij (na primere konnektorov so znacheniem neposredstvennogo sledovanija [Cognitive-semantic approach to the description of means

of connection of sentences (by the example of connectors with the meaning of direct following)]. Trudy Instituta russkogo jazyka im. V. V. Vinogradova. 2016. No. 10. Pp. 120–133.

14. *Kobozeva I.* (2017). Konnektory kontaktnogo predshestvovanija v russkom i frantsuzskom jazykakh v zerkale nadkorpusnoj bazy dannykh [The connectors expressing contact precedence in Russian and French as reflected in the supracorpora database]. Contrastive linguistics. 2017. Vol. XLII, No. 4. Pp. 48–62.
15. *MAS* — Slovar' russkogo yazyka [Dictionary of the Russian language]: In 4 vol. A. P. Evgen'eva (ed.), Moscow, Russkii Yazyk, 1981.
16. *Morkovkin V. V. ed.* (1997). Slovar' strukturnykh slov russkogo yazyka [Dictionary of structural words of the Russian language], Moscow, Lazur'.
17. *Mustaioki A., Kopotev M.* (2004). K voprosu o statuse ekvivalentov slova tipa *potomu chto, v zavisimosti ot, k sozhaleniyu* [On the status of word-equivalents of the type *potomu chto, v zavisimosti ot, k sozhaleniyu*]. Voprosy jazykoznanija. 2004. No. 3. Pp. 88–107.
18. *Nikolaeva T. M.* (1985). Funktsii chastits v vyskazyvanii [Functions of particles in the utterance]. Moskva, Nauka.
19. *Nikolaeva T. M.* (2008). Neparadigmaticheskaya lingvistika (istoriya «bluzhdayushchikh chastits») [Non-paradigmatic linguistics (history of the "wandering particles")]. Moscow, Yazyki Slavyanskikh Kul'tur.
20. *NKRYa* — Natsional'nyi korpus russkogo yazyka [Russian National Corpus], available at: www.ruscorpora.ru.
21. *Oates S.* (2000). Multiple Discourse Marker Occurrence: Creating Hierarchies for Natural Language Generation, Proceedings of ANLP-NAACL 2000, April 29—May 4, 2000. Seattle, Washington, USA. Pp. 41–45.
22. *Prasad R., Bunt H.* (2016). Semantic relations in discourse: the current state of ISO 24617-8. H. Bunt (ed.), Proceedings og 12th Joint ACL-ISO Workshop on Interoperable semantic annotation, 28 May 2016, Slovenia, Portorož. Pp. 80–92. Available at: http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-ISA12proceedings.pdf.
23. *Priyatkina A. F.* (1977). Ob otlichii soyuza ot drugikh svyazuyushchikh sredstv [On the difference between the conjunction and other connecting words]. Russkii yazyk v shkole. 1977. No. 4. Pp. 102–106.
24. *Rogozhnikova R. P.* (2003). Tolkovyi slovar' sochetanii, ekvivalentnykh slovu [Explanatory dictionary of constructions equivalent to the word], Moscow, Astrel'-AST.
25. The Penn Discourse Treebank, available at: https://www.seas.upenn.edu/~pdtb/.
26. *Webber B. L., Joshi A. K.* (1998). Anchoring a Lexicalised Tree-Adjoining Grammar for Discourse. M. Stede, L. Warner, E. Hovy (eds), Discourse Relations and Discourse Markers. Proceedings from the Workshop. COLING-ALC'98, Montreal. Pp. 86–92.
27. *Webber B. L.* (2016). Concurrent Discourse Relations, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016", Moscow, June 1–4, 2016. available at: http://www.dialog-21.ru/media/3488/webber.pdf.

НАСКОЛЬКО ЛИНГВОСПЕЦИФИЧЕН СОЮЗ *ХОТЯ*?[1]

Инькова О. Ю. (Olga.Inkova@unige.ch)
ИПИ ФИЦ ИУ РАН, Москва, Россия;
Женевский университет, Женева, Швейцария

Нуриев В. А. (nurieff.v@gmail.com)
Институт языкознания РАН, Москва, Россия;
ИПИ ФИЦ ИУ РАН, Москва, Россия

TO WHAT EXTENT IS THE CONJUNCTION *KHOTYA* LANGUAGE-SPECIFIC?

Inkova O. Yu. (Olga.Inkova@unige.ch)
Institute of Informatics Problems, FRC CSC RAS, Moscow, Russia; University of Geneva, Geneva, Switzerland

Nuriev V. A. (nurieff.v@gmail.com)
Institute of Linguistics, RAS, Moscow, Russia; Institute of Informatics Problems, FRC CSC RAS, Moscow, Russia

The paper describes the Russian connective *khotya* ('although') from a contrastive perspective. First, it focuses on the semantic description of the connective and proposes to differentiate its four meanings, namely, concessive propositional, concessive illocutionary, adversative propositional and adversative illocutionary. The paper analyzes the functioning of the connective *khotya* (prototypical marker of concessive relations) and that of the connective *no* ('but', prototypical marker of adversative relations). In so doing, it comes to the following conclusion: the adversative meaning of *khotya* develops on the basis of its concessive meaning as the connection between the situations presented in the textual fragments that are linked by the connective becomes less logical. Similarly, i. e. vice-versa, as the logical connection between situations becomes stronger, this gives rise to a concessive interpretation in utterances with *no*. Further, the paper takes a closer look at French equivalents *khotya* gets, when occurring in each of its four meanings. The concluding section attempts to define the degree of language-specificity of *khotya*. To this end, several parameters

[1] Исследование выполнено в ИПИ ФИЦ ИУ РАН в рамках проекта по гранту РНФ № 16-18-10004.

are considered: (1) cases where the connective has a zero equivalent, (2) cases of divergent translation (the connective is translated by a non-connective), (3) number of translation patterns. To perform a contrastive analysis and to collect statistical data, the supracorpora database of connectives is used. The database is built upon the parallel Russian-French and French-Russian subcorpora of the RNC.

Keywords: supracorpora database of connectives, Russian, French, contrastive linguistics, corpus linguistics, language-specific connectives

Союз *хотя* имеет сложную семантическую организацию и обладает разветвленной конфигурацией значений. В этой связи закономерно возникает вопрос, насколько эта конфигурация является лингвоспецифичной. В большинстве своем исследования, посвященные *хотя*, ориентируются на русскоязычный материал, лишь в некоторых случаях используются отдельные сопоставительные наблюдения [см., например, В. Апресян 2006]. Однако, как известно, сопоставление часто позволяет увидеть то, что остается незамеченным при моноязычном подходе, а также понять, насколько изучаемая языковая единица уникальна по своим функциональным свойствам. Чтобы определить степень лингвоспецифичности *хотя* (для анализа мы взяли только «одиночное» *хотя*), мы сравним его с его словарным эквивалентом во французском языке *bien que* и применим критерии для определения лингвоспецифичности коннекторов, предложенные в работе [Инькова 2017]. Для сопоставительного анализа и получения статистики использована надкорпусная база данных коннекторов (НБД), созданная на основе русско-французского и французско-русского подкорпусов Национального корпуса русского языка (НКРЯ)[2]. Однако, прежде чем перейти непосредственно к анализу, обратимся к семантике *хотя*.

1. Семантика *хотя*: существующие описания и некоторые уточнения

Союзу *хотя* посвящена обширная литература (см., например, В. Апресян 2006, 2015, Богомолова 1955, Гречишникова 1971, Николаева, Фужерон 1999, Печенкина 1976, Прияткина 1968, Санников 2008, Теремова 1986, Урысон 2002, 2003, 2011 и др.). Неоднократно предпринимались попытки его лексикографического описания, в частности, в словарях служебных языковых единиц [например, Бурцева 2010, Ефремова 2004]. Как в лингвистических исследованиях, так и в лексикографических источниках у союза *хотя* выделяются два основных значения: уступительное и противительное, которые, несмотря на разные теоретические подходы, описываются одинаково. Уступительное *хотя* выражает онтологическую несовместимость между двумя связываемыми

[2] Принцип устройства и система аннотирования в НБД описаны в работе [Inkova, Popkova 2017]; фрагмент НБД доступен по адресу: http://a179.ipi.ac.ru/PublicLingvoProjects/main.aspx.

им ситуациями (1), а противительное *хотя* — противоположность оценок одной и той же ситуации или предмета (2).

(1) Глядя на других, Илья Ильич и сам перепугался, *хотя* и он и все прочие знали, что начальник ограничится замечанием. [И. А. Гончаров. Обломов (1859)]

(2) Обломов был проще Штольца и добрее его, *хотя* не смешил ее так или смешил собой и так легко прощал насмешки. [И. А. Гончаров. Обломов (1859)]

В (1) имеется ситуация *p = Илья Ильич и сам перепугался* и ситуация *q = все знали, что начальник ограничится замечанием*. Как правило, ситуация *q* сопровождается ситуацией не-*p*, т. е. в нашем примере: если начальник ограничивается замечанием, то пугаться *не* надо. Отношения между не-*p* и *q* можно, следовательно, подвести под некоторую закономерность, которую нарушает реализация *p*. В (2) такую — логическую — зависимость между *p* и *q* установить невозможно. Кроме того, *p* и *q* описывают не две несовместимые ситуации, а две характеристики одного человека, первая из которых (*p = Обломов был проще Штольца и добрее его*) на основе наших знаний об устройстве мира может считаться положительной, а вторая (*q = не смешил ее так*...) — отрицательной. Но между свойствами *p* и *q* нет онтологического противоречия (они вполне могут быть представлены как имеющие одинаковую оценку; ср. *Обломов был проще Штольца и добрее его и не смешил ее так*). При помощи *хотя* говорящий просто заставляет слушающего пересмотреть выводы, которые он мог сделать на основании *p*.

В РГ-80 (§ 3054), а также в работе [В. Апресян 2010] *хотя* уступительное и *хотя* противительное считаются двумя лексемами, семантически никак не связанными между собой[3], хотя в РГ-80 (§ 3052) и подчеркивается, что «[п]ротивительное значение является обязательным компонентом любой разновидности уступительных отношений». В пользу того, что уступительное и противительное значение нельзя рассматривать как «две ступени градации одного и того же значения противопоставления (с большей или меньшей степенью противоречивости соотносимых частей)», в РГ-80 приводятся следующие примеры:

(3) а. Прости меня, *но* ты живешь отсталыми представлениями о природе авторства (П. Крон);

б. Мне могут не поверить, *но* бывают случаи, когда, даря свои книги, авторы, мягко говоря, намекают: не пора ли, мол, и на меня написать пародию. (газ.)

[3] Противоположной точки зрения придерживается Е. Урысон [Урысон 2003], которая предлагает рассматривать *хотя1* и *хотя2* как неточные конверсивы; в словаре [Морковкин 1997] у *хотя* выделяются три значения, названия которых (уступительное, уступительно-противительное и противительно-уступительное) отражают ‘плавный’ переход от уступительности к противительности.

(4) Она несла в руках отвратительные, тревожные желтые цветы. Черт их знает, как их зовут, *но* они первые почему-то появляются в Москве (Булг.); То ли память о молодости цепка, то ли ход мыслей такой, *но* всякий раз размышления о жизни приводят в село. (Шукш.)

В этих высказываниях, как справедливо замечает РГ-80 (§ 3054), действительно нельзя заменить *но* на *хотя*, но дело здесь не в «степени противоречивости частей», а в том, на каком уровне *но* устанавливает логико-семантическое отношение (ЛСО) «вопреки ожиданиям», один из видов противопоставления. Как было показано в [Инькова-Манзотти 2001], *но* может устанавливать это ЛСО как на пропозициональном уровне, т. е. между положениями вещей, описанными в каждом из связываемых им фрагментов текста, так и на уровне речевых актов. В последнем случае различаются две группы употреблений. В первой из них *q* противопоставляется *p* как акту речи, имеющему определенную иллокутивную цель: речевой акт *q* осуществляется несмотря на его несоответствие целевой установке, заданной речевым актом *p*, или в нарушение условий его успешности.

(5) Прости, что писал тебе злые письма, *но* в Шахматове всегда обо всем беспокоишься больше, чем где бы то ни было. (А. Блок; пример из Инькова-Манзотти 2001: 223)

В (5) *p* является речевым актом извинения, но нарушается одно из условий его успешности: условие искренности говорящего (когда человек просит прощения, предполагается, что он признает свою вину). Однако в *q* говорящий сводит извинения на нет, и все высказывание воспринимается скорее как оправдание: «Я прошу прощения за злые письма, но я не считаю себя виноватым, поскольку по объективным причинам я не мог поступить иначе».

Вторую группу высказываний с союзом *но*, устанавливающим ЛСО «вопреки ожиданиям» на уровне акта речи, составляют случаи, затрагивающие постулаты речевого общения, сформулированные Г. П. Грайсом, а также общие правила коммуникации — этические, социальные и др., выступающие в роли коммуникативных регуляторов. В таких случаях речевой акт *q* осуществляется, несмотря на нарушение определенного постулата или правила, о чем сигнализирует *p*. Так, в (3а) выше *p* сигнализирует о том, что будет нарушен принцип вежливости, а в (3б) — о нарушении правила качества информации. В (4) в *p* говорящий предупреждает слушающего о том, что не располагает информацией в полном объеме: в а. — о причине положения вещей, описанного в *q*, в б. — об одном из аспектов описываемого в *q* предмета.

Схожие употребления имеет и *хотя*[4], как в уступительном (6), так и в противительном значениях (7).

[4] Возможность употребления, которое квалифицируется вслед за [Иорданская, Мельчук 2007: 430] как «иллокутивное», т. е. когда союз характеризует отношение между пропозициональным смыслом одного предложения и фактом произнесения другого предложения, допускается для *хотя* в работе [В. Апресян 2015: 30], но не анализируется, а механизм его функционирования не раскрывается.

(6) — Белорус не годится. Белорусов навалом. Узбека мне давай, или, на худой конец, эстонца… *Хотя* нет, погоди, эстонец вроде бы есть… [Сергей Довлатов. Чемодан (1986)]

(7) Вглядываясь, вдумываясь в свой быт и все более и более обживаясь в нем, он наконец решил, что ему некуда больше идти, нечего искать, что идеал его жизни осуществился, *хотя* без поэзии <…>. [И. А. Гончаров. Обломов. (1848–1859)]

При выражении уступительных ЛСО на уровне речевых актов, *хотя* вводит такой речевой акт, в большинстве случаев, ассертивный, который перечеркивает иллокутивную силу речевого акта, осуществленного при помощи *p*. В (6) в *p* говорящий формулирует просьбу, которая, хоть и частично (та часть, которая касается эстонца), но аннулируется в *q* (ср. *нет*, которое следует за *хотя*).

При противительных ЛСО, устанавливаемых на уровне речевого акта, *p* всегда будет ассертивным речевым актом, сообщающим некоторую информацию, один из аспектов которой уточняется в *q*, и *q* таким образом ограничивает истинность сказанного в *p* (ср. термин «ограничение», используемый в БАС для определение этого значения *хотя*, «оговорка» у В. Апресян 2006, «поправка» в Морковкин 1997). В (7) выше истинность утверждения *идеал его жизни осуществился* пересматривается, ‘оговаривается’: осуществиться он осуществился, но не полностью, а *без поэзии*.

Как и при противительных ЛСО, устанавливаемых между двумя пропозициями, при противительных ЛСО, устанавливаемых на уровне речевых актов, высказывание описывает одну ситуацию. Но сначала взятую в целом (в *p*), а затем (в *q*) в одном из ее аспектов. На структурном уровне такая семантическая конфигурация проявляется в синтаксической зависимости слова или словосочетания, вводимого *хотя*, от сказуемого *p*, как в (7), которое может повторяться в *q* (8):

(8) Здесь с самого начала все предрешено, закрыто и **завершено**, *хотя*, *правда*, **завершено** не в одной плоскости. [М. М. Бахтин. Проблемы поэтики Достоевского. (1963)]

Возвращаясь к употреблению *хотя* на пропозициональном уровне, можно отметить следующее: чтобы понять различия в функционировании *хотя* и *но*, необходимо учитывать критерий наличия / отсутствия прямой логической зависимости между компонентами *p* и *q*. При отсутствии логической зависимости, т. е. при невозможности подвести отношения между *p* и *q* под некоторую общую закономерность, замена *но* на *хотя* возможна, но, в отличие от предложений с уступительными отношениями, невозможно изменить порядок следования *p* и *q*:

(9) Балабанову удалось понять то, чего не понял Бодров. *Но* страшно неровная картина (*МК-бульвар*, пример из Инькова-Манзотти 2001: 175)
Балабанову удалось понять то, чего не понял Бодров. *Хотя* страшно неровная картина

??*Хотя* страшно неровная картина, Балабанову удалось понять то, чего не понял Бодров.

При наличии логической зависимости между *p* и *q* необходимо различать две конфигурации. В первой из них противопоставление между *p* и *q* устанавливается не напрямую, а через имплицитные звенья *r* и не-*r*: *p* → *r* ***но*** *q* → не-*r*[5].

(10) Известно, что жизнь Ивана Грозного унесла комета. Царь-чудовище хотел казнить астрологов, предсказавших по ней день его смерти, *но* не успел. Умер. (*Новые Известия*, пример из Инькова-Манзотти 2001: 185)

P = Иван Грозный хотел казнить астрологов аргументирует или, по крайней мере, создает ожидание, в пользу вывода *r = Астрологи были казнены*; *q = Не успел. Умер* аргументирует в пользу противоположного вывода не-*r = Астрологи не были казнены*. Такая зависимость между *p* и *q* называется «косвенной». Замена *но* на *хотя* невозможна.

Если же *p* и *q* соотносятся без участия имплицитного вывода не-*r*, точнее этот вывод совпадает с *q*, то такая зависимость считается «прямой».

(11) Они говорят о нём, болване, портаче, а потом не только о нём, <…> затем уже совсем о других, о таких, о которых говорить не полагается, *но* они *всё равно* говорят. [Ю. О. Домбровский. Факультет ненужных вещей, часть 2 (1978)]

В (11) *p = о таких вещах говорить не полагается* аргументирует в пользу вывода *r = о них не будут говорить*, это вывод опровергается непосредственно в *q* (= не-*r*) = *они все равно говорят*. При прямой логической зависимости между *p* и *q* семантический механизм функционирования *но* совпадает с семантическим механизмом, лежащим в основе уступительного *хотя*, но в зеркальном отражении, отсюда необходимость при замене *хотя* на *но* и наоборот изменить порядок следования фрагментов текста:

p ***хотя*** *q* (*q* → не-*p*)
p → *r* ***но*** *q* (*q* = не-*r*), т. е. *p* → не-*q* ***но*** *q*
Они об этом все время говорят, *хотя* об этом говорить не принято.
Об этом говорить не принято, *но* они все время об этом говорят.

Таким образом, если на пропозициональном уровне уступительное значение у *но* возникает при усилении логической связи между *p* и *q*, то можно предположить, что противительное значение *хотя* развивается, наоборот, в тех случаях, когда логическая связь между *p* и *q* ослабевает, и они воспринимаются как разнонаправленные аргументы.

[5] На эту особенность в свое время указал [Ю. И. Левин, 1970], однако сравнивая функционирование *но* и *хотя*, он этот параметр не использует.

2. Сопоставительный анализ

Для сопоставительного анализа были взяты 259 случаев употребления одиночного *хотя* в функции коннектора, аннотированных в НБД. Т.е. не берутся случаи сочетания *хотя* с частицами *и* и *бы*, а также двухместные коннекторы, в состав которых он может входить. Сразу заметим, что в 207 случаях (т.е. в 80%) часть, вводимая *хотя*, находится в постпозиции, в 13 случаях — в интерпозиции или во вставочном предложении в скобках. Преобладающим значением одиночного *хотя* является уступительное пропозициональное — 226 случаев (87,3%), далее со значительным отрывом следуют уступительное «иллокутивное», т.е. установленное на уровне речевого акта, — 15 случаев (5,8%), противительное пропозициональное — 11 (4,2%) и противительное иллокутивное — 7 (2,7%). В каждом из этих четырех значений *хотя* имеет свои функциональные особенности, и ему соответствуют во французском языке разные эквиваленты.

2.1. Уступительные пропозициональные ЛСО

Когда *хотя* реализуется в данном значении, фрагмент текста, вводимый им, имеет свободную позицию, но с преобладанием постпозиции: 175 случаев из 220; в 11 случаях он находится в интерпозиции или во вставочном предложении в скобках. Все случаи с препозицией фрагмента текста, вводимого *хотя*, приходятся на это его значение. При переводе на французский язык пост- и интерпозиция фрагмента текста, вводимого *хотя*, в 8 случаях заменена на препозицию:

(12) Я тотчас узнал его, *хотя* он весь закутался в темный плащ и шляпу надвинул на лицо. [И. С. Тургенев. Первая любовь. (1860)]
Bien qu'il fût entièrement enveloppé dans un manteau noir et eût enfoncé son chapeau sur les yeux je le reconnus immédiatement [Trad. par M.-R. Hoffman. (1974)]

(13) Женя, *хотя* не собиралась делать этого, рассказала Лимонову о своих делах с пропиской. [В. С. Гроссман. Жизнь и судьба. (1959)]
Bien qu'elle n'eût pas l'intention de le faire, elle raconta toute son histoire à Limonov. [Trad. par A. Berelowitch. (1980)]

Наиболее частотной моделью перевода для уступительного *хотя* является *bien que* (50,4%), также прототипический показатель этих ЛСО во французском языке. Следующим по частотности идет нулевой эквивалент (9,3%). Но в большинстве случаев это связано либо с тем, что фрагмент текста с *хотя* не переведен (14), либо с тем, что он вводит однородное сказуемое в отрицательной форме. Французский язык в таких случаях предпочитает использовать предлог *sans*, у которого некоторые исследователи видят в отдельных случаях уступительное значение [Morel 1996: 85–89], как в приведенном примере, где ситуации исключают друг друга. На наш взгляд, основное значение этого

предлога — отрицание сопутствующего обстоятельства, а значит, он переводит скорее отрицание однородного сказуемого в приведенном примере.

(14) Ей всегда казалось, что он лучше всех в семье понимает ее, *хотя* он мало говорил о ней. [Л. Н. Толстой. Анна Каренина (1873–1877)]
Il lui semblait toujours que de toute la famille nul ne la comprenait mieux que son père. [Trad. par H. Mongault. (1952)]

(15) Оба смеемся, *хотя* не говорим ничего смешного.
[А. П. Чехов. Скучная история. (1889)]
Nous rions tous deux, *sans* avoir rien dit de drôle.
[Trad. par E. Parayre. (1960)]

Среди других французских соответствий преобладают языковые единицы, выражающие уступительные ЛСО (в совокупности около 16 %): *même si, malgré, quoique, encore que, nonobstant, quelque… que, si… n’en… pas moins*; или близкое к нему ЛСО «вопреки ожиданиям» (13,72 %): *pourtant, cependant, mais*. *Хотя* может быть переведен и глаголами (*ne pas empêcher, avoir beau*) или глагольными формами с уступительным значением (деепричастие).

Заметим, что в 3,10 % случаев уступительное *хотя* переводится показателем сопоставительного ЛСО *alors que*.

(16) И выбрался через окно. *Хотя* вполне мог пройти через дверь.
[С. Д. Довлатов. Чемодан. (1986)]
Et il enjamba la fenêtre. *Alors qu*’il aurait pu sortir par la porte.
[Trad. par J. Michaut-Paternò. (2001)]

По-видимому, переводчик не видит здесь сильной логической зависимости между *p* и *q*, а лишь некоторое несоответствие между ними, что подтверждает нашу гипотезу о близости уступительного и противительного значений *хотя*.

2.2. Уступительные иллокутивные ЛСО

В отличие от уступительного пропозиционального значения, которое реализуется, как правило, в рамках сложного предложения (192 случая, т. е. 89,3 %), уступительное иллокутивное значение *хотя* реализуется в большинстве случаев между двумя самостоятельными высказываниями, разделенными точкой: в нашем корпусе 13 из 15. В двух других случаях, *q* может рассматриваться как вставка с функцией комментария *p* (17), или же *p* и *q* принадлежат разным говорящим: прямая речь и слова автора (18). Порядок следования *p* и *q* — фиксированный: *q*, вводимое *хотя*, всегда в постпозиции:

(17) А коли дома читать будешь, *хотя* на это надежды мало, в тетрадочку конспектируй… [Аркадий Вайнер, Георгий Вайнер. Эра милосердия. (1975)]
Si, toutefois, tu parviens à les lire chez toi, ce dont je doute fort, tu devras en faire le résumé dans un cahier… [Переводчик неизвестен. (2005)]

(18) — Как хорошо, — *хотя* ничего хорошего в дранке, висевшей с потолка, в куче штукатурки в углу, в безобразной трубе не было. [В. С. Гроссман. Жизнь и судьба. (1960)]
«C'est bien!» *bien qu*'il n'y eût rien de bien ni dans le tuyau de cheminée difforme, ni dans les débris qui pendaient du plafond, ni dans le tas de plâtre dans un coin. [Trad. par A. Berelowitch. (1980)]

Пример (18) интересно сравнить с (19), чтобы понять различие между иллокутивным и пропозициональным уступительным значением *хотя*:

(19) — Раз живой, то, значит, порядок, — давясь пылью, кашляя и отхаркивая, сказал он, *хотя* порядка было не так уж много. [В. С. Гроссман. Жизнь и судьба. (1960)]
— Alors, tout va bien, dit-il en toussant et crachant, *bien que* les choses n'allassent pas si bien que cela. [Trad. par A. Berelowitch. (1980)]

В (19) речевое действие названо глаголом *сказать*, и отношение устанавливается между двумя ситуациями *p = сказал, что порядок* и *q = порядка не было*, тогда как в (18) *хотя* непосредственно связывает речевое действие *p* (факт произнесения высказывания) с ситуацией, в которой оно производится (с пропозициональным содержанием *q*).

При переводе *хотя* в этом значении используются единицы, которые могут выражать ЛСО на уровне высказывания. Наиболее частотным (26,67%) является наречие *quoique*, на долю *bien que* приходится всего 20% (ср. с 50% с уступительным пропозициональным), поскольку, являясь подчинительным союзом, он здесь находится на пределе своих функциональных возможностей. 13,3% приходится на долю *d'ailleurs* 'впрочем', которое, как и его русский эквивалент, выражает противительное ЛСО на уровне речевого акта [Инькова 2013]. Нулевой эквивалент, как в (17), зарегистрирован в 6,67% случаев.

2.3. Противительные пропозициональные и противительные иллокутивные ЛСО

Статистические данные в НБД по противительным отношениям, выражаемым *хотя*, пока не являются представительными, но сформулируем тем не менее некоторые наши наблюдения. Противительное пропозициональное значение *хотя* реализуется как в рамках сложного предложения (7 случаев из 11), так и между самостоятельными предложениями (5 из 11). Порядок следования фрагментов текста — фиксированный, всегда с постпозицией *q*. Противительное иллокутивное *хотя* реализуется в рамках предложения, что объясняется лежащей в его основе семантической структурой (см. примеры выше), также с фиксированным порядком, исключающим препозицию *q*, которое находится либо в постпозиции (6 случаев из 7), либо имеет вставочный характер: введя тему своего высказывания, говорящий затем как бы антиципирует свою поправку в то, что им будет сказано, придав ей меньшую значимость.

(20) Тут прикованность к миру, *хотя* в отрицательной форме, остается полностью. [Н. А. Бердяев. О самоубийстве. (1931)]

Несмотря на то, что *bien que* при переводе *хотя* в противительном пропозициональном значении отличается продуктивностью (на него приходится наибольшее число случаев — 27,27%), следующими по частотности идут показатели ЛСО «вопреки ожиданиям» *cependant* и *mais*, на долю которых в совокупности приходится 36,4%. Остальные эквиваленты (*ceci dit, et, même si,* нулевой) зафиксированы единичными употреблениями.

(21) Так прошел сентябрь, наступила осень. *Хотя* на газонах еще зеленела трава и днем было жарко, как в мае… [С. Д. Довлатов. Иностранка. (1986)]
Ainsi passa septembre, l'automne arriva. *Cependant* l'herbe des pelouses était encore verte et dans la journée il faisait aussi chaud qu'en mai…
[Trad. par J. Michaut-Paternò. (2001)]

Наиболее частотной моделью перевода иллокутивного противительного *хотя* является в НБД *bien que* (43%). Заметим, однако, что переводчик каждый раз восстанавливает в *q* полную предикативную структуру, поскольку для *bien que* в меньшей степени характерно вводить член предложения; ср перевод (7) выше:

(22) <…> l'idéal de sa vie s'était réalisé, *bien que* ce fût sans poésie <…>
[Trad. par L. Jurgenson. (1988)]

Более естественными в таких структурах будут *mais* (28,6%) и *quoique* (14,3%).

(23) Буран еще продолжался, *хотя* с меньшею силою.
[А. С. Пушкин. Капитанская дочка. (1836)]
Le bourane durait encore, *mais* avec une moindre violence.
[Trad. par L. Viardot (1853)]

(24) Это будет исповедь, как Голядкин, *хотя* в другом тоне и роде
[М. М. Бахтин. Проблемы поэтики Достоевского (1963)]
Ce sera une confession, comme Goliadkine, *quoique* dans un autre genre et un autre ton [Trad. par Kolitcheff. (1970)]

3. Заключительные замечания: насколько лингвоспецифичен союз *хотя*?

В [Инькова 2017] были предложены следующие критерии для количественного анализа лингвоспецифичности коннекторов: 1) число случаев перевода нулевым эквивалентом; 2) число случаев дивергентного перевода (переводится не коннектором); 3) число моделей перевода. На основе данных о моделях перевода, зафиксированных для каждого коннектора, НБД программно высчитывает рейтинг моделей перевода — относительную величину, которая тем

выше, чем больше у коннектора зафиксировано моделей перевода[6]. Согласно статистике, представленной в **Таблице 1**, *хотя* находится в зоне средней степени лингвоспецифичности. Если за минимум взять *короче,* для которого рейтинг моделей перевода составляет 0,29, а за максимум принять *вообще* (1,90), то *хотя* со своим 0,89 находится в зоне средней степени лингвоспецифичности. *Хотя* нечасто переводится нулевым эквивалентом или не коннектором. Тем не менее на долю *bien que*, словарного эквивалента *хотя*, приходится меньше половины случаев перевода (47,13%).

Таблица 1. Лингвоспецифичность *хотя*: показатели НБД

Коннектор	Всего	Нулевой эквивалент	Дивергентный перевод	Модели перевода	Рейтинг моделей перевода
вообще	163	29 (17,79%)	11 (6,75%)	52	1,90
хотя	261	24 (9,2%)	12 (4,6%)	33	0,89
короче	58	3 (5,17%)	2 (3,45%)	4	0,29

В пользу лингвоспецифичности *хотя* говорят также данные французско-русского подкорпуса НКРЯ: выборка параллельных контекстов, сформированная по поисковому запросу «хотя (CONJ)», из 76 случаев «одиночного» употребления *хотя* в переводах на русский в 22 (29%) *хотя* не имеет стимульных лексических единиц в оригинале, в 17 (22%) стимулом перевода послужило *bien que*, в 13 (17%) — *quoique*, в 8 (11%) — *mais*. По два раза (3%) встречаются: (*et*) *pourtant*, (*et*) *cependant*, *sans*, *et*. Единичные случаи: *alors que*, *malgré*, *même si*, *quand*, *sauf que*, *sauf à*, *si*, *toutefois*. При этом нельзя сказать, что добавление *хотя* при переводе с французского является индивидуальной переводческой стратегией, поскольку оно фиксируется в переводах О. де Бальзака, Ф. Бегбедера, Ж. Верна, В. Гюго, П. Модиано у разных переводчиков.

Литература

1. *Apresjan V. Ju.* (2006), Linguistic concession [Ustupitel'nost' v yazyke], Linguistic worldview and systematic lexicography [Jazykovaja kartina mira i sistemnaja leskikografija], Ju. D. Apresjan (ed.), Jazyki slavjanskih kul'tur, Moscow, pp. 615–712.
2. *Apresjan V. Ju.* (2010), Examples of the dictionary entries: 'Khotja' [Obraztsy slovarnykh statej: 'Khotja'], Prospectus of the Active dictionary of Russian [Prospekt aktivnyogo slovarja russkogo jazyka], Ju. D. Apresjan (ed.), Jazyki slavjanskih kul'tur, Moscow, pp. 199–200.
3. *Apresjan V. Ju.* (2015), Concession: the mechanisms of formation and interaction of complex meanings in the language [Ustupitelnost': mekhanismy obrazovanija

[6] Рейтинг учитывает тот факт, что прирост моделей перевода не является линейным и рассчитывается по формуле: $R_i = n_i/(m_i)^{0,65}$, где $i = 1, \ldots, L$ (общее число аннотаций с данным коннектором в НБД). Подробнее см. [Inkova & Kruzhkov in press].

i vzaimodejstvija slozhnykh znachenij v jazyke], Jazyki slavjanskoj kul'tury, Moscow.
4. *BAS* (1950–1965) — Dictionary of modern standard Russian: in 17 volumes [Slovar' sovremennogo russkogo literaturnogo jazyka: V 17 tomakh], Nauka, Moscow-Leningrad.
5. *Bogomolova A. V.* (1955), Concessive constructions with the conjunction khotja (khot') in modern standard Russian [Ustupitel'nye konstruktsii s sojuzom khotja (khot') v sovremennom russkom literaturnom jazyke], Synopsis of the PHD thesis, Leningrad.
6. *Burtseva V. V. ed.* (2010), Dictionary of the adverbs and structural words of Russian language [Slovar' narechij i sluzhebnyh slov russkogo jazyka], Drofa, Moscow.
7. *Efremova T. F.* (2004), Explanatory dictionary of the Russian auxiliary parts of speech [Tolkovyj slovar' sluzhebnyh chastei rechi russkogo yazyka], Astrel'-Ast, Moscow.
8. *Grechishnikova R. M.* (1971), A complex sentence with phraseological means conveying concessive relations in modern standard Russian [Slozhnoe predlozhenie s frazeologizirujushhimisja sredstvami vyrazhenija ustupitel'nyh otnoshenij v sovremennom literaturnom russkom jazyke], Synopsis of the PHD thesis, Leningrad.
9. *Inkova-Manzotti O. Ju.* (2001), Adversative connectives in French and Russian [Konnektory protivopostavlenija vo frantsuzskom i russkom jazykakh], Informelektro, Moscow.
10. *Inkova O.* (2013), "As for the rest, madam the Marquise…": about the semantics of *v ostal'nom, v pročem and vpročem* ["A v ostal'nom, prekrasnaja markiza…": o semantike *v ostal'nom, v prochem i vprochem*], Russian language and linguistic theory [Russkij jazyk v nauchnom osveschenii], Vol. 25, No 1, pp. 21–40.
11. *Inkova O.* (2017), Principles of How to Determine the Degree of Language-specificity of Connectives [Printsipy opredelenija stepeni lingvospetsifichnosti konnektorov], Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference "Dialog" [Komp'juternaja Lingvistika i Intellektual'nye Tekhnologii: Trudy Ezhegodnoj Mezhdunarodnoj Konferentsii "Dialog"], RSUH, Moscow, Vol. 2, pp. 150–160.
12. *Inkova O., Kruzhkov M.* (in press), Statistical analysis of language specificity of connectives based on parallel texts, Proceeding of International conference on statistical analysis of textual data JADT-2018, Rome, 12–15 June 2018.
13. *Inkova O., Popkova N.* (2017), Statistical data as information source for linguistic analysis of Russian connectors, Inform. Appl. [Informatika i ee primeneniya], Vol. 11, No. 3., pp. 123–131.
14. *Iordanskaja L. N., Mel'chuk I. A.* (2007), The sense and the combinatory in dictionary [Smysl i cochetaemost' v slovare], Jazyki slavjanskoj kul'tury, Moscow.
15. *Levin Ju. I.* (1970), On one group of Russian conjunctions [Ob odnoj gruppe sojuzov russkogo jazyka], Machine Translation and Applied Linguistics [Mashinny Perevod i Prikladnaja Lingvistika], Moscow, Vol. 13, pp. 64–88.
16. *MAS* (1981) — Dictionary of the Russian language [Slovar' russkogo yazyka]: In 4 vol. A. P. Evgen'eva (ed.), Russkij Jazyk, Moscow.

17. *Morel M.-A.* (1996), La concession en français, Ophrys, Paris.
18. *Morkovkin V. V. ed.* (1997), Dictionary of structural words of the Russian language [Slovar' strukturnyh slov russkogo jazyka], Lazur', Moscow.
19. *Nikolajeva T. M., Fougeron I.* (1999), Some observations on semantics and the status of complex sentences with concessive conjunctions [Nekotorye nabljudenija nad semantikoj i statusom slozhnyh predlozhenij s ustupitel'nymi sojuzami], Issues in linguistics [Voprosy jazykoznanija], No 1, pp. 17–36.
20. *Pechenkina T. G.* (1976), Syntactic category of concession and forms of its expression in the standard Russian language of the 2nd half of the 19th century [Sintaksicheskaja kategorija ustupitel'nosti i formy ee vyrazhenija v russkom literaturnom jazyke vtoroj poloviny XIX veka], Synopsis of the PHD thesis, Leningrad.
21. *Prijatkina A. F.* (1968), On the conjunction KHOTJA in modern Russian [O sojuze HOTJA v sovremennom russkom jazyke], The Russian language at school [Russkij jazyk v shkole], No 2, pp. 83–88.
22. *RG-80* — Russian grammar (1980), [Russkaja grammatika], N. Ju. Shvedova (ed.), Vol. 2, Nauka, Moscow.
23. *Sannikov V. Z.* (2008), Russian syntax into semantic-pragmatic space, Jazyki slavjanskih kul'tur, Moscow.
24. *Teremova R. M.* (1986), Concessive semantics and its expression in modern Russian [Semantika ustupitel'nosti i ee vyrazhenie v sovremennom russkom jazyke], A. I. Herzen Leningrad State Pedagogical Institute, Leningrad.
25. *Uryson E. V.* (2002), The conjunction KHOTJA through the prism of semantic primitives [Sojuz HOTJA skvoz' prizmu semanticheskih primitivov], Issues in linguistics [Voprosy jazykoznanija], No 6, pp. 35–54.
26. *Uryson E. V.* (2003), Semantic and valency structure of the words with concessive meaning [Semanticheskaja i valentnaja struktura slov s ustupitel'nym znacheniem], Russian language and linguistic theory [Russkij jazyk v nauchnom osveschenii], Vol. 6, No 2, pp. 217–246.
27. *Uryson E. V.* (2011), Describing the semantics of conjunctions: Linguistic data on the activities of consciousness [Opyt opisanija semantiki sojuzov: Lingvisticheskie dannye o dejatel'nosti soznanija], Jazyki slavjanskih kul'tur, Moscow.

ЕЩЕ РАЗ О МИКРОКОНСТРУКЦИЯХ, СФОРМИРОВАННЫХ СЛУЖЕБНЫМИ СЛОВАМИ: *ТО И ДЕЛО*[1]

Иомдин Л. Л. (iomdin@iitp.ru)

Институт проблем передачи информации им. А. А. Харкевича РАН; Российский государственный гуманитарный университет; Москва, Россия

Статья продолжает серию исследований микросинтаксиса русского языка, которые автор проводит на протяжении достаточно продолжительного времени. В центре внимания находится адвербиальная микросинтаксическая единица *то и дело*, которая представляется весьма интересной и поучительной, поскольку сочетает в себе ряд имплицитных семантических особенностей и уникальный набор синтаксических свойств, часть из которых обнаруживается благодаря рассмотрению не только синхронных, но и диахронных языковых данных. Эта единица исследуется на фоне других микросинтаксических элементов, которые оказываются ее соседями по словарю, но обладают существенно другим набором лингвистически релевантных свойств. Обсуждаются вопросы, связанные с адекватным представлением фразеологических единиц типа *то и дело* в Микросинтаксическом словаре русского языка, составляемом автором и его коллегами, и в корпусе текстов, содержащем микросинтаксическую аннотацию.

Ключевые слова: микросинтаксис, синтаксические фраземы, Микросинтаксический словарь русского языка, корпус с микросинтаксической разметкой

ONCE AGAIN ON MICROSYNTACTIC CONSTRUCTIONS FORMED WITH FUNCTIONAL WORDS: *TO I DELO* ‘EVERY NOW AND THEN’

Iomdin L. L. (iomdin@iitp.ru)

Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Moscow, Russia

[1] Работа выполнена при частичной поддержке Российского научного фонда (грант № 16-18-10422). Автор выражает фонду искреннюю признательность.

The paper continues a series of research studies into the microsyntax of Russian, conducted by the author for a considerable period of time. Specifically, the focus is on the adverbial syntactic idiom *to i delo* '≈ every now and then', which seems very interesting and instructive as it combines implicit semantic features and a unique set of syntactic facets that could be revealed by both present-day and diachronic linguistic data. This syntactic idiom is considered against the background of other microsyntactic elements that happen to be its neighbors in the dictionary but feature a substantially different set of linguistically relevant properties. It is shown how phraseological units of such kind can be presented in the Microsyntactic dictionary of Russian, under development by the author and his colleagues, and in the corpus of texts annotated with microsyntactic phenomena.

Keywords: Microsyntax, syntactic idioms, Microsyntactic dictionary of Russian, corpus annotated with microsyntactic elements

1. Вводные замечания. Современное состояние Микросинтаксического словаря

Предлагаемая статья продолжает широкое исследование микросинтаксических единиц русского языка, в течение полутора десятилетий проводимое автором и несколькими его коллегами в Лаборатории компьютерной лингвистики Института проблем передачи информации им. А. А. Харкевича РАН. В рамках данного исследования был опубликован ряд статей (см., например, [Апресян, Иомдин 1990], [Иомдин, 2003, 2014, 2015, 2017а, 2017б], [Iomdin, 2005], [Маракасова, Иомдин, 2016].

К настоящему времени создан представительный прототип Микросинтаксического словаря русского языка, насчитывающий 1000 единиц. В этот словарь входят лексикографические единицы двух типов: нестандартные синтаксические конструкции и синтаксические фраземы.

Оба эти типа единиц хорошо известны в лингвистической литературе (хотя часто и под другими названиями, в частности, фразеосхемы, фразеомодели, синтаксические идиомы или синтаксически нечленимые конструкции; см., например, [Шведова, 1958], [Шмелев, 1976], [Jackendoff, 1997], [Меликян, 2004], [Mel'čuk, 2012], [Apresjan V., 2014]), однако ни тот, ни другой до сих пор не получали исчерпывающего и даже систематического описания. Причина состоит в том, что мы имеем здесь дело с областью языковых явлений, находящихся на перекрестье грамматики и словаря. Из-за этого промежуточного положения данная область оказывается обделенной вниманием лингвистов-теоретиков и лексикографов.

Следует сказать, что микросинтаксический подход к описанию языковых явлений, используемый автором, идейно близок к грамматике конструкций, хотя и отличается от него в некоторых важных отношениях. В частности, он затрагивает не весь синтаксис языка, а лишь его периферийную часть, которая противопоставляется базовому «большому» синтаксису. Поэтому целиком полагаться на грамматику конструкций в решении указанных задач мы не решаемся — иначе есть риск утратить различие между нашими

единицами и неидиоматичными базовыми конструкциями языка, с одной стороны, и разницу между этим конструкциями и не имеющими синтаксического своеобразия фразеологическими выражениями (типа *метать бисер перед свиньями* или *не хватать звёзд с неба*), с другой стороны.

Сколько-нибудь полного инвентаря микросинтаксических единиц какого бы то ни было языка до сих пор не существовало. Это и понятно, так как такие единицы слишком индивидуальны, чтобы массово появляться в грамматиках. Но они слишком индивидуальны и для словаря: обычная лексикографическая практика традиционных словарей состоит в том, что в словарных статьях одного или более слов, входящих в какую-либо микросинтаксическую единицу, последние упоминаются и в лучшем случае снабжаются краткими пояснениями. Между тем весьма часто такие единицы заслуживают полноценной словарной статьи, а то и — в случае многозначности — нескольких.

Одной из основных целей создания Микросинтаксического словаря было хотя бы частично заполнить лакуну в языковом описании, которая соответствует микросинтаксису.

Нестандартных синтаксических конструкций, входящих в Микросинтаксический словарь, немного (всего порядка трех десятков).

Примерами таких конструкций являются, с одной стороны, конструкции так называемого малого синтаксиса, такие как инфинитивно-модальные конструкции типа *Y-у X-овать* (‘Y должен будет X-овать’, ср. *Мне сегодня всю ночь работать*), *Y-у не X-овать* (‘Отсутствует перспектива, что Y будет X-овать’, ср. *Тебе никогда не выиграть у этого мастера по шахматам*), конструкции типа *Y-у не до X-а* (‘Y занят более важными делами, чем X, и заявляет, что не будет делать X, считая, что X-ом можно пренебречь’; ср. *Ему сейчас не до развлечений*). Примечательно, что у двух последних из трех упомянутых конструкций существуют варианты вообще без отрицательной частицы: *Разве тебе выиграть у этого мастера по шахматам?*; *До развлечений ли ему сейчас?*

Другие примеры нестандартных синтаксических конструкций — разнообразные конструкции с повторяющимися или коллокативно повторяющимися лексическими единицами и всякий раз своеобразной семантикой; ср. *Видеть не видел, но кое-что слышал* (≈ ‘не видел, но имел место факт «слышал», что является чем-то более слабым, чем факт «видел»*); Жалуйся, не жалуйся, результата не будет* (≈ ‘Независимо от того, пожалуешься ты или нет, это не приведет к результату…’*); Мальчишки есть мальчишки* (≈ ‘Мальчишки ведут себя так, как следует ожидать от мальчишек’); *Машина как машина, ничего особенного* (≈ ‘Ничем не примечательная машина’*), Служба службой, — хмуро сказал он, — а здоровье здоровьем* (≈ ‘Разумеется, службу надо исполнять хорошо, но и здоровьем нельзя пренебрегать’ — М. Ибрагимбеков), *спать сном праведника* (≈ ‘спать так крепко, как, в предположении говорящего, спят праведники*), пасть смертью героя* (≈‘погибнуть как герой’) и т. д.

К нестандартным синтаксическим конструкциям относятся также разговорные единицы типа (а) *Как ты?,* (б) *Я хорошо <вполне нормально, неплохо, так себе>.* , на первый взгляд очень близкие к конструкциям типа (в) *Мне было хорошо,* (г) *Кому тут грустно?,* (д) *Нам всем очень жаль,* (е) *Мальчику было*

очень больно и т. п. Последние конструкции (типа «в–е») состоят из предикатива в качестве сказуемого (возможно, с нулевой или эксплицитной связкой) и субъекта, выраженного дательным падежом и, вообще говоря, являются достаточно стандартными, хорошо известными и всесторонне исследованными. Между тем в конструкциях типа «а–б» падеж субъекта выражен дефолтным именным падежом, и такие конструкции не получили, по нашему мнению, достаточного отражения ни в теоретическом синтаксисе, ни в лексикографии. Эти конструкции тем не менее заслуживают внимания, так как обладают заметной синтаксической и семантической спецификой. В частности, синтаксическая специфика заключается в том, что в качестве предиката функционируют бесспорные наречия, а семантика таких предложений достаточно четко очерчена: как вопрос, так и ответ касаются физического состояния или жизненных обстоятельств отвечающего собеседника или кого-то (или чего-то), принадлежащего к его личной сфере. Трудно представить себе серьезный обмен репликами между московскими пенсионерами типа *Вы не знаете, как там император Акихито? — Да вроде он отлично.* Эти конструкции введены в научный оборот в самое последнее время и тоже должны включаться в Микросинтаксический словарь.

Основную часть словника Микросинтаксического словаря составляют синтаксические фраземы — такие фразеологические единицы, которые, помимо лексической избирательности и семантической некомпозициональности, обладают заметной синтаксической спецификой.

Некоторые такие единицы тяготеют к цельным словам; ср. единицы типа *все равно* (по крайней мере в некоторых из значений), *все же, тот же, между тем, между прочим, тем не менее, тем более, только что, пока что, разве что, то и дело,* частицы *как бы* или *что ли* и сотни других. В процессе языковых изменений такие единицы движутся по направлению к единому слову подобно тому, как это произошло с союзом/наречием *якобы* (которое, согласно данным НКРЯ, могло писаться раздельно вплоть до начала XX века) или подобно партикулам [Т. М. Николаевой 2008], таким как *либо* и *неужели*, исторически образовавшимся из нескольких отдельных формантов.

Прочие же единицы к цельным нечленимым словам свести нельзя или же это весьма затруднительно сделать (ср. *как быть* — в примерах типа *как быть преподавателям со студентами, которые пользуются шпаргалками, то ли дело* — как в примере *Он не любит делать уроки, то ли дело мультики смотреть; черт <дьявол, бес, бог…>, знает кто <где, когда, зачем, …>.*

Значительное количество статей Микросинтаксического словаря посвящено составным предлогам типа *в адрес, в течение, во время, в отличие от, в связи с, на тему, по причине, в отсутствие* и многим другим. В принципе эти единицы, которые разумно считать синтаксическими фраземами, не представляют особой сложности при описании, но часто обладают нетривиальными свойствами, игнорируемыми в традиционных словарях. Например, сложный предлог *на тему* допускает заполнение своей валентности в виде местоименного прилагательного типа *этот, какой,* стоящего перед именной частью фраземы и согласующегося с ним (ср. *на эту тему, на какую тему*), но практически не допускает такое заполнение притяжательными местоимениями типа *мой*

или *ваш*. Предлог *в связи с* может выступать в варианте типа *в этой связи, в какой связи,* «обрубая» последний элемент *с,* в то время как трехчастный предлог *в отличие от* ничего такого не допускает, ср. *в отличие от этого*, но не **в это отличие*). Предлог *насчёт* в каноническом случае не представляет микросинтаксической единицы, будучи отдельным словом, но образует синтаксическую фразему в случаях типа *на мой счет, на этот счет на сей счет, на Петин счет, на чей счет* — ср. *Княжна меня решительно ненавидит; мне уже пересказали две-три эпиграммы на мой счёт, довольно колкие, но вместе очень лестные* (М. Ю. Лермонтов) и т. *д.*

Некоторые языковые элементы приобретают статус микросинтаксических единиц в результате орфографических капризов языка.

Например, *все-таки* — единое слово и не входит в число микросинтаксических единиц. Зато *всё же* и *всё ж таки* таковыми единицами являются.

В некоторых случаях приходится решать сложный вопрос о статусе единицы: слово это или не слово. Так, *то-то же* и *то-то и оно* — без сомнения, неоднословные единицы, а как поступать с изолированным элементом *то-то* (как, например, в предложении *То-то я смотрю, он такой злой ходит*. — В. Войнович.)?

Значительные трудности представляет собой разграничение в словаре между близкими микроединицами. Например, может показаться, что в выражениях (а) *в чём дело и* (б) *в чём же дело* представлена одна и та же единица, однако скорее всего это не так: (б) может означать «не вижу препятствий», а (а), пожалуй, такого значения не имеет.

Составление Микросинтаксического словаря проходит параллельно с созданием корпуса текстов, содержащих микросинтаксическую разметку. При выборе корпуса для такой разметки оказалось разумным использовать уже существующий глубоко аннотированный корпус русских текстов СинТагРус (см., в частности, [Дяченко и др., 2015]). Мы исходили из предположения, что чем глубже уровень разметки текста, тем качественнее и надежнее трактовка микроединиц. Уже первые результаты микросинтаксической разметки показали, что микросинтаксические единицы — весьма частое языковое явление. Даже по грубым оценкам, не менее 25 процентов предложений длиной в 12 слов и более содержат хотя бы одну микросинтаксическую единицу. Некоторые из них удается идентифицировать именно благодаря исследованию корпуса.

Сегодня СинТагРус содержит свыше 1 млн. словоупотреблений (около 70 тыс. предложений, входящих приблизительно в семьсот русских текстов нескольких языковых жанров). Однако уже сейчас ясно, что этого корпуса недостаточно для поддержки и развития Микросинтаксического словаря: многие единицы в СинТагРус’е попросту отсутствуют или представлены единичными вхождениями. Предстоит создать новый корпус, специально ориентированный на микросинтаксис. В настоящей статье вопрос о создании нового корпуса с микросинтаксической разметкой мы оставляем в стороне.

2. Синтаксическая фразема «то и дело»

Очень многие единицы Микросинтаксического словаря представляют не только лексикографический, но и общенаучный интерес. Одна из таких единиц — это достаточно нетривиальная адвербиальная синтаксическая фразема *то и дело,* к описанию которой мы и переходим. Эта синтаксическая фразема выступает в конструкциях типа

(1) *Мы то и дело сверялись с картой* (Д. Гранин. Зубр).

Как уже отмечалось в одной из предыдущих работ автора на тему микросинтаксической фразеологии [см. Иомдин, 2017а], статьи Микросинтаксического словаря составляются по единой схеме, в значительной мере опирающейся на опыт создания нового [Активного словаря русского языка под ред. Ю. Д. Апресяна 2014]. Материал статьи распределяется по восьми основным зонам:

а) имя микросинтаксической единицы;
б) тип микросинтаксической единицы (синтаксическая фразема или нестандартная синтаксическая конструкция);
в) лексический состав единицы (конкретные слова синтаксической фраземы или классы слов, входящих в состав нестандартной синтаксической конструкции, если таковые имеются);
г) лексикографическое толкование единицы (по возможности, аналитическое);
д) модель управления (если она имеется);
е) синтаксические свойства единицы, формулируемые в терминах грамматики зависимостей и включающие идентификацию синтаксических отношений между элементами единицы и синтаксических отношений, встраивающих данную единицу в предложение;
ж) синонимы и аналоги;
з) комментарии;
и) иллюстрации.

Эта схема является максимальной и применяется к наиболее сложным микросинтаксическим единицам. В других случаях задействуется лишь часть указанных зон, однако зоны а), б), в), ж) и з) являются обязательными.

Синтаксическую фразему *то и дело* мы будем представлять в виде словарной статьи Микросинтаксического словаря, задействуя релевантные зоны. В ходе изложения будут приводиться комментарии, которые, строго говоря, не являются частью словарной статьи, но необходимы для лучшего понимания материала.

(а) Имя: ТО И ДЕЛО

(б) Тип: синтаксическая фразема

(в) Лексический состав: ТО1, И2, ДЕЛО1

(г) Синтаксические свойства. Обязательные элементы фраземы — местоименное существительное ТО1 (в именительном падеже), ограничительная

частица И2 и существительное ДЕЛО1 в именительном падеже единственного числа. Вершиной конструкции является слово ДЕЛО1, от него по предикативному синтаксическому отношению подчиняется местоименное существительное ТО1, а по ограничительному отношению частица И2. Местоименное существительное ТО1 является представителем синтаксической фраземы во внешних связях: оно синтаксически подчиняет по эксплетивному синтаксическому отношению предикатное слово, которое заполняет единственную семантическую валентность фраземы (в приведенном выше примере это глагол *сверялись*: см. синтаксическую структуру на рис. 1 ниже).

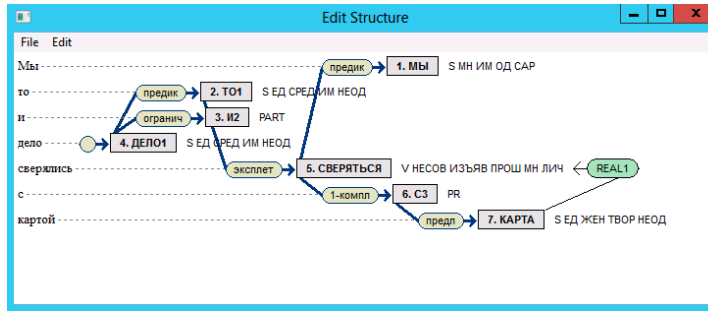


Рис. 1. Теоретически обоснованная синтаксическая структура предложения (1) ***Мы то и дело сверялись с картой.***

(д) Лексикографическое толкование: *То и дело P:*

Имеет место ряд ситуаций или событий Р. Р повторяется многократно. Говорящий наблюдает за Р и считает, что 1) количество повторений событий Р в период наблюдения велико2; 2) повторы случаются более или менее равномерно; 3) эти повторы происходит «сами по себе», без целенаправленного воздействия со стороны участников Р.

(е) Модель управления.

В семантической структуре единицы *то и дело* имеется переменная, соответствующая единственной (пассивной) валентности ситуации Р. Эта валентность должна быть насыщена. В большинстве случаев Р выражается глаголом в любой форме (личной, инфинитивной, причастной или деепричастной), ср.

(2) *В их речи то и дело звучали* [Р] *строки, фразы, стихи.* [Д. Гранин. Зубр];

(3) *В технологии приготовления кумыса есть одна обязательная операция: молоко необходимо то и дело взбалтывать* [Р] — *несколько тысяч раз, как кто-то подсчитал.* [«Наука и жизнь», 2008];

[2] Образно говоря, говорящий ощущает эти повторения как происходящие непрерывно: как ни глянешь, событие Р имеет место. Это можно сравнить с явлением, известным в науке о зрении как критическая частота мельканий (critical flicker frequency).

(4) *В тумане маячила отчуждённая фигура рыбака, то и дело махающего* [Р] *удилищем.* [В. Астафьев. Затеси];

(5) *Ведь художник пишет самого себя, то и дело поглядывая* [Р] *в зеркало.* [А. Рекемчук. Мамонты].

Помимо этого, валентность Р может выражаться любым словом, которое может выступать в качестве сказуемого: прилагательным, предикативом, предлогом во главе предложно-именной группы; ср.

(6) *То и дело слышен смех* [Ю. Казаков. Голубое и зеленое],

(7) *А нам то и дело стыдно за него.* [Александр Володин. Одноместный трамвай],

(8) *Умные то и дело на работе, а дурак завсегда на печи валяется* [Записи русских народных сказок А. Н. Афанасьева].

Здесь требуется дать важные пояснения. Синтаксическая структура, предложенная в синтаксической зоне словарной статьи (г), обусловлена диахронической картиной развития данной синтаксической фраземы. Как мы убедились в ходе исследования материалов корпусов русского языка, до самого конца XIX века единица *то и дело* была способна управлять придаточным, вводимым союзом *что*, которое и вводило валентность Р. Ср. следующие примеры из НКРЯ:

(9) *Собою я, говорят, хорош, танцую весьма порядочно, да и в обществе я довольно ловок: в мазурке у Армидиных меня то и дело что выбирают…* [В. А. Соллогуб. Большой свет, 1840];

(10) *Хозяйка то и дело что кланялась дорогим гостям, прося не побрезгать скромным угощением, чем Бог послал* [В. А. Соллогуб. Тарантас, 1845];

(11) *На душе повеселело / Вдруг у мужиков:/ По рукам же то и дело / Что гуляет штоф.* [Л. Н. Трефолев. Обоз, 1864];

(12) *Ему во всяком случае будет не хорошо: царь Николай не любит, чтобы русские заживались в Париже, а он то и дело, что в Париже.* [П. П. Вяземский. Письма и записки Оммер де Гелль, 1887];

(13) *Он не забыл своей собственной тяжелой юности, когда окружающие то и дело что подрывались под него, стараясь лишить его наследия отца.* [А. К. Шеллер-Михайлов. Дворец и монастырь, 1900].

Параллельно с этой синтаксической фраземой использовались (а отчасти используются до сих пор) две другие, очень близкие к рассматриваемой как по смыслу, так и по синтаксису: *Только и делает, что…* и *То и делает, что,* ср следующие примеры из НКРЯ*:*

(14) *В тот день я только и делал, что ее фотографировал* [А. Снегирев];

(15) *Всю свою жизнь он только и делал, что общался с беременными.* [Т. Соломатина].

(16) *Да он только то и делает, что стоит около меня и рассказывает мне, как все будет хорошо.* [В. Шахиджанян].

Конструкция *то и делает* почти всегда сопровождается ограничителем типа *только* или *лишь,* хотя иногда выступает и изолированно:

(17) *Местное население то и делает, что пьет, следовательно, молодым студентам нечего делать вечером* [с сайта фанатов Любови Полищук]

В этих конструкциях, как представляется автору, наблюдается типичное явление эксплетива, выражаемого местоименным существительным *то*, которое берет на себя роль номинализатора. Кажется, что именно это явление имеет место и в рассматриваемой нами синтаксической фраземе *то и дело.* Правда, в синтаксических фраземах, представленных в примерах (14–16), фигурирует не существительное *дело,* а глагол *делать,* и бесспорного перехода *делает Þ дело* в истории развития русского языка нам обнаружить не удалось. Тем не менее кажется все-таки естественным предположить, что синтаксическое устройство этих вариантов фразем тождественно: *то* — это эксплетив, подчиняющий союз *что*.

Частичным подтверждением такого вывода можно считать длительное параллельное существование в русском языке конструкций, где *что* может выступать при глаголе (примеры (18) и (21)), а может и при существительном, этот глагол номинализующем (примеры (19–20) и (22): во всех таких случаях этот союз подчиняется эксплетиву *то* или, неожиданным образом, слову *только*:

(18) *С утра он только и думает, что об обеде, а после обеда об ужине* [А. А. Потехин. Вакантное место, 1870]

(19) *Только и мысль одна, что об них…* [А. А. Потехин. Шуба овечья — душа человечья, 1854];

(20) *У него в голове только и мыслей, что про походы; сердце его только и бьется для военной славы.* [П. А. Кулиш. Черная рада, 1846–1857];

(21) *Все только и говорили, что о докладе Хрущева.* [И. Эренбург. Люди, годы, жизнь];

(22) *Вечером только и разговоров было, что о дебюте «полковника» в кино.* [В. Аксенов. Пора, мой друг, пора].

(все примеры из НКРЯ, обращает на себя внимание использование обеих конструкций одним и тем же автором в примерах (18) и (19)).

Любопытно, что и слово *дело* в таких конструкциях может стоять в родительном падеже, ср.

(23) *С утра до вечера, то и дела, что сидит в шинке!* [Н. В. Гоголь. Сорочинская ярмарка];

(24) *Будто у меня только и дела, что помнить о галошах.* [М. Сергеев. Волшебная галоша].

Если союз *что* при *то и дело* отсутствует, ситуация не меняется — подчиненным членом эксплетива оказывается глагол (или другой предикат, как бы оставшийся после удаления *что*). Тем самым наша фразема функционирует в составе конструкций с вложенным придаточным.

Добавим, что в этом отношении конструкция *то и дело* не одинока. Близкие по синтаксису ситуации имеют место в конструкциях с союзом *благо*, допускающим как варианты без *что* так и со *что*; ср.

(25) *Благо профессорская гостиная позволяла рассадить полста людей* (Ю. Трифонов. Дом на набережной, 1976);

(26) *Благо что до него было метров двести* [А. Иванов. Географ глобус пропил, 2002]

и в конструкциях с предикативом *хорошо* (которое можно считать зарождающимся союзом, ср. (Иомдин, 2014]):

(27) а. *Хорошо меня там не было* vs.
б. *Хорошо что меня там не было*).

Следует особо подчеркнуть, что такая трактовка рассматриваемой синтаксической фраземы, включающая эксплетивное то, является сугубо научной. В Микросинтаксическом словаре использовать ее кажется целесообразным.

В то же время для компьютерно-лингвистических задач, в том числе при микроскопической разметке корпуса, следует применять упрощенную синтаксическую трактовку, допустимую в предположении, что мы ограничиваемся рассмотрением только синхронного состояния языка. (Одним из оснований для упрощения является тот факт, что при изложенной научной трактовке синтаксическая структура предложения с *то и дело* окажется непроективной, хотя очевидного нарушения дефолтного порядка слов тут не наблюдается).

Здесь возможны (и, по-видимому, одинаково приемлемы на практике) два упрощенных варианта синтаксической трактовки фраземы: (1) интерпретация формы *то* как местоименного прилагательного ТОТ в среднем роде и именительном падеже и подчинение его существительному ДЕЛО1 по определительному синтаксическому отношению; частица И2, как и в научном варианте, подчиняется существительному ДЕЛО1 по ограничительному отношению); (2) интерпретация всей фраземы как единого нерасчлененного слова — наречия, т. е.своего рода безусловного оборота). И в том, и в другом случае результирующая синтаксическая структура содержащего эту фразему предложения останется проективной. Оба варианта структуры представлены на рис. 2 и 3 соответственно.

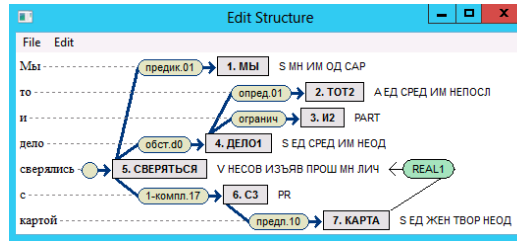


Рис. 2. Упрощенная синтаксическая структура предложения (1)
Мы то и дело сверялись с картой.

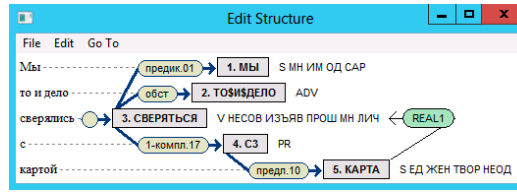


Рис. 3. Синтаксическая структура предложения (1)
Мы то и дело сверялись с картой. при представлении синтаксической фраземы в виде единого узла (безусловного оборота)

Стоит заметить, что в отсутствие научной трактовки нам не удалось бы продемонстрировать глубокие синтаксические различия между *то и дело* и другими микросинтаксическими единицами, по существу сформированными из тех же словарных единиц. Так, например, единица *то и дело* синтаксически организована совсем не так, как соседствующая с этой единицей в словаре синтаксическая фразема *то ли дело*; ср.

(28) *То ли дело рюмка рома, / Ночью сон, поутру чай; / То ли дело, братцы, дома!.. / Ну, пошел же, погоняй!..* [А. С. Пушкин. Дорожные жалобы];

(29) *То ли дело вскарабкаться на скалу и оттуда с вытянутыми вперед руками броситься вниз головой в быструю воду* [В. П. Беляев. Старая крепость].

Единица *то ли дело* выступает в (28) и (29) (а, вероятно, и во всех других случаях, как субстантивное сказуемое при подлежащем (которое может быть именной группой или инфинитивом)). Внутри этого сказуемого *то,* бесспорно, является местоименным прилагательным, являющимся определением при *дело* — т. е. именно тем элементом, который мы постулировали в упрощенном понимании синтаксиса *то и дело.*

Дополнительным доказательством адекватности этой трактовки *то ли дело* являются конструкции, являющиеся фактически антонимами этой единицы — *другое дело, иное дело* (часто выступающее в паре с *одно дело*) и даже *не то дело*, ср.

(30) *Но одно дело прочесть об этом в пьесе Шекспира, а другое дело показать в кино* [А. Сокуров];

(31) *Ноябрь будет подличать и ненавидеть. Не то дело июль. Июль высокий и живой* [Н. Садур].

И в (30), и в (31) имеет место определительная связь между *дело* и прилагательным.

(ж) Комментарии.

1) Единица то и дело не допускает при себе ни отрицания, ни квантификации: нельзя сказать

(32) **Поезда не то и дело опаздывают*

или

(33) **Поезда весьма <слишком, очень, не очень> то и дело опаздывают*.

Сюда не относится достаточно искусственная, но в общем стандартная ситуация, когда отрицание используется в качестве возражения к употребленной номинации, как в следующем диалоге:

(34) А. *Поезда то и дело опаздывают.* — Б. *Они опаздывают не «то и дело», а лишь иногда.*

2) Практически невозможно употребить *то и дело* в вопросе:

(35) **То и дело ли поезда опаздывают?*

(36) ??*Разве поезда то и дело опаздывают?*

3) Отрицание при выражении, реализующем валентность ситуации Р, напротив, допускается, хотя частотность такого отрицания весьма ограничена; ср.

(37) *Его истерическая натура то и дело не выдерживала, и он срывался на рыдания* [Дмитрий Липскеров. Последний сон разума].

4) Неуместно использовать синтаксическую фразему *то и дело* в ситуациях, оцениваемых как нормальные (пусть даже в реальной жизни такие ситуации не редкость):

(38) #*Поезда то и дело приходят вовремя,*

(39) #*Вася то и дело водит в садик младшую сестру,*

(40) ?*Ева то и дело помогает бабушке печь блины.*

С другой стороны, оценка ситуации не как нормальной, а как хорошей использованию синтаксической фраземы *то и дело* не мешает. Вполне характерный пример — это пастернаковская строфа (см. ниже пример 46).

Заметим, впрочем, что из 15 присутствующих в размеченном микросинтаксическими единицами корпусе СинТагРус фраз с *то и дело* (см. рис. 4) лишь одну (фразу 11) можно оценить как описывающую хорошую ситуацию:

(41) *Однако они то и дело приглашали нас из детсада поразвлекать их концертом.*

При ближайшем рассмотрении, впрочем, выясняется, что они — это живущие впроголодь в блокадном Ленинграде моряки, которые волею случая оказались вне своего корабля.

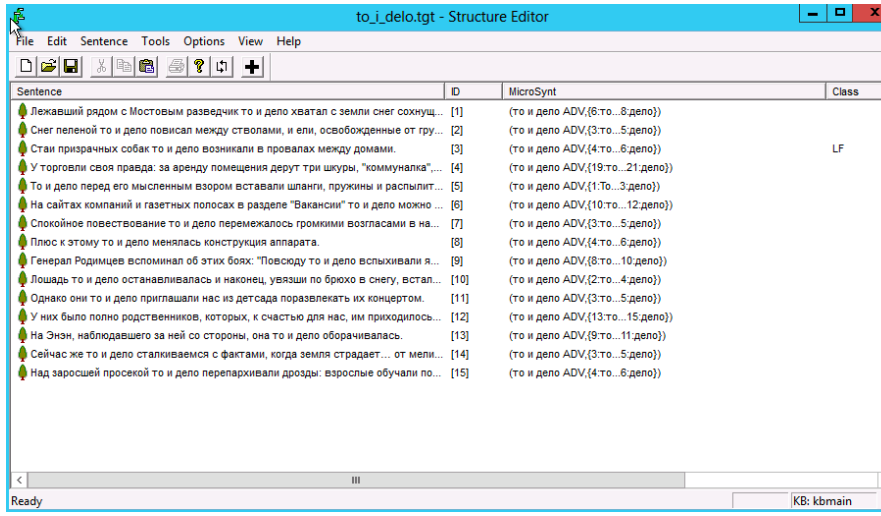


Рис. 4. Фрагмент микросинтаксически размеченного корпуса СинТагРус, содержащий ситаксическую фразему *то и дело*. Фрагмент не содержит ни одного ложно-положительного вхождения этой единицы, что свидетельствует о ее фразеологической силе.

А из 30 фраз с *то и дело*, первыми полученных поиском по НКРЯ, около 25 описывают плохие ситуации, несколько фраз описывают нейтральные ситуации и лишь одна — однозначно хорошую:

(42) *Читали, вчитывались, запоминали, цитировали. В их речи то и дело звучали строки, фразы, стихи* [Д. Гранин. Зубр].

(3) Иллюстрации. Синтаксическая фразема *то и дело* достаточно частотна и встречается в текстах любых жанров, за исключением, пожалуй, научно-технического.

Примеры из корпуса «СинТагРус»:

(43) *Снег пеленой то и дело повисал между стволами, и ели, освобожденные от груза, раскачивали лапами* [художественная проза: Ю. Казаков. Двое в декабре];

(44) *То и дело перед его мысленным взором вставали шланги, пружины и распылители воды* [научно-популярная литература, Наука и жизнь № 8, 2004].

Примеры из основного корпуса НКРЯ см. выше.
Примеры из поэтического корпуса НКРЯ:

(45) *Жил-был славный царь Дадон. / С молоду был грозен он / И соседям то и дело / Наносил обиды смело; Но под старость захотел / Отдохнуть от ратных дел* [А. С. Пушкин. Сказка о золотом петушке];

(46) *Мело весь месяц в феврале, / И то и дело / Свеча горела на столе, / Свеча горела.* [Б. Пастернак. Зимняя ночь].

Литература

1. *Активный словарь русского языка* (2014). Т I. Авторы: Апресян В. Ю., Апресян Ю. Д., Бабаева Е. Э., Богуславская О. Ю., Галактионова И. В., Гловинская М. Я., Иомдин Б. Л., Крылова Т. В., Левонтина И. Б., Лопухина А. А., Птенцова А. В., Санников А. В., Урысон Е. В. Под ред. акад. Ю. Д. Апресяна. М.: Языки славянской культуры. 408 с.
2. *Апресян Ю. Д., Иомдин Л. Л.* (1990). Конструкции типа НЕГДЕ СПАТЬ в русском языке: синтаксис и семантика // Семиотика и информатика. М.; Вып. 29. С. 3–89.
3. *Дяченко П. В., Иомдин Л. Л., Лазурский А. В., Митюшин Л. Г., Подлесская О. Ю., Сизов В. Г., Фролова Т. И. , Цинман Л. Л.* (2015). Современное состояние глубоко аннотированного корпуса текстов русского языка (СинТагРус) // Национальный корпус русского языка. 10 лет проекту. Труды Института русского языка им. В. В. Виноградова. М. Вып. 6. С. 272–299.
4. *Иомдин Л. Л.* (2003). Большие проблемы малого синтаксиса // Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям Диалог'2003. Протвино, 2003. С. 216–222.
5. *Иомдин Л. Л.* (2013). ЧИТАТЬ НЕ ЧИТАЛ, НО…: об одной русской конструкции с повторяющимися словесными элементами // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог» (2013). М.: Изд-во РГГУ. Вып. 12. Т. 1. С. 272–284.
6. *Иомдин Л. Л.* (2014). Хорошо меня там не было: синтаксис и семантика одного класса русских разговорных конструкций // Сб. статей «Grammaticalization and Lexicalization in the Slavic Languages». По материалам Международного симпозиума «Грамматикализация и лексикализация в славянских языках», 11–14 ноября 2011 г. München-Berlin-Washington/D.C.: Verlag Otto Sagner. Band 55. P. 423–436.
7. *Иомдин Л. Л.* (2015). Конструкции микросинтаксиса, образованные русской лексемой раз // SLAVIA, časopis pro slovanskou filologii, ročník 84, 2015, sešit 3, s. 291–306.
8. *Иомдин Л. Л.*(2017a). Как нам быть с конструкциями типа как быть? // Компьютерная лингвистика и интеллектуальные технологии:

По материалам ежегодной Международной конференции «Диалог» (Москва, 31 мая —3 июня 2017 г.). М.: Изд-во РГГУ, 2017. Вып. 16 (23). Т. 1. С. 161–176.

9. *Л. Л. Иомдин* (2017б). Между синтаксической фраземой и синтаксической конструкцией. Нетривиальные случаи микросинтаксической неоднозначности. SLAVIA, časopis pro slovanskou filologii, ročník 68, 2017, sešit 2–3, s. 230–243.
10. *Маракасова А. А., Иомдин Л. Л.* (2016). Микросинтаксическая разметка в корпусе русских текстов СинТагРус // Информационные технологии и системы 2016 (ИТиС’2016). Сборник трудов 40-ой междисциплинарной школы-конференции ИППИ РАН. Репино, Санкт-Петербург. С. 445–449.
11. *Меликян В. Ю.* (2004). Современный русский язык. Синтаксис нечленимого предложения: Учебное пособие. Ростов-на-Дону: Изд-во РГПУ.– 288 с.
12. *Николаева Т. М.* (2008). Непарадигматическая лингвистика (История «блуждающих частиц»). М.: Языки славянских культур. — 689 с.
13. *Шведова Н. Ю.* (1958). О некоторых типах фразеологизированных конструкций в строе русской разговорной речи // Вопр. языкозн., № 2, С. 95–100.
14. *Шмелев Д. Н.* (1976). Синтаксическая членимость высказывания в современном русском языке. М., Наука, 1976. 152 с.
15. *Apresjan, Valentina* (2014). Syntactic idioms across languages: corpus evidence from Russian and English. // Russian Linguistics, Vol. 38, Issue 2, pp 187–203.
16. *Iomdin, Leonid* (2005). A Hypothesis of Two Syntactic Starts // Восток — Запад: Вторая международная конференция по модели «Смысл — Текст». Отв. ред. Ю. Д. Апресян, Л. Л. Иомдин. М.: «Языки славянской культуры». С. 165–175.
17. *Iomdin, Leonid* (2016). Microsyntactic Phenomena as a Computational Linguistics Issue. // Grammar and Lexicon: Interactions and Interfaces. Proceedings of the Workshop. Osaka, Japan. 2016, pp. 8–18. (http://aclweb.org/anthology/W/W16/W16-38.pdf). ISBN 978-4-87974-706-8
18. *Jackendoff, Ray* (1997). Twisting the Night Away. // Language, Vol. 73, pp. 534–559.
19. *Mel’čuk, Igor* (2012). Phraseology in the language, in the dictionary, and in the computer. // Yearbook of Phraseology, Volume 3, Issue 1, pp. 31–56.

References

1. *Active dictionary of Russian. [Aktivnyj slovar' russkogo jazyka].* (2014). Vol I. Authors: Apresjan V. Yu., Apresjan Yu. D., Babaeva E. E., Boguslavskaya O. Yu., Galaktionova I. V., Glovinskaya M. Ya., Iomdin B. L., Krylova T. V., Levontina I. B., Lopukhina A. A., Ptentsova A. V., Sannikov A. V., Uryson E. V. Edited by Yu. D. Apresjan. Moscow: Jazyki slvjanskoj kultury. 408 p. (In Russian).
2. *Apresjan Juri, Iomdin Leonid.* (1990). Constructions like NEGDE SPAT' ('There is nowhere to sleep') in Russian: Syntax and Semantics // Semiotika i informatika. Moscow, Issue 29. P. 3–89. (In Russian).
3. *Apresjan, Valentina* (2014). Syntactic idioms across languages: corpus evidence from Russian and English. // Russian Linguistics, Vol. 38, Issue 2, pp 187–203.
4. *Djachenko P. V., Iomdin L. L., Lazursky A. V., Mitjushin L. G., Podlesskaja O. Yu., Sizov V. G., Frolova T. I., Tsinman L. L.* (2015). The state-of-the-art of a deeply annotated corpus of Russian texts (SynTagRus). [Sovremennoe sostojanie gluboko annotirovannogo korpusa tekstov russkogo jazyka (SynTagRus)] // Nacionalnyj korpus russkogo jazyka. 10 let proektu. Trydy Instituta russkogo jazyka im. V. V. Vinogradova. Moscow. Vyp. 6. p. 272–299. (In Russian).
5. *Iomdin, Leonid* (2003). Big Problems of Minor Syntax. [Bol'šie problem malogo sintaksisa] // Kompjuternaja lingvistika I intellektual'nye tekhnologii. Po materialam ezhegodnoj Mezhdunarodnoj konferencii «Dialog» (2003). Protvino, 2003. Pp. 216–222. (In Russian.)
6. *Iomdin, Leonid* (2005). A Hypothesis of Two Syntactic Starts // East-West. 2nd International conference on Meaning — Text Model. Juri Apresjan, Leonid Iomdin (eds.) Jazyki slavjanskoj kultury P. 165–175. (In Russian.)
7. *Iomdin L. L.* (2013). CHITAT' NE CHITAL, NO: On a Russian construction with repeated word elements. [CHITAT' NE CHITAL, NO: ob odnoj russkoj konstrukcii s povtorjajushchimisja slovesnymi elementami] // Kompjuternaja lingvistika I intellektual'nye tekhnologii. Po materialam ezhegodnoj Mezhdunarodnoj konferencii «Dialog» (2013). Moscow: Izdatelstvo RGGU. Issue 12. Vol 1. p. 272–284. (In Russian).
8. *Iomdin L. L.* (2014). Good thing I wasn't there: Syntax and Semantics of a class of Russian colloquial constructions. [Khorosho menja tam ne bylo: sintaksis i semantika odnogo klassa russkikh razgovornykh konstrukcij] // In: Grammaticalization and Lexicalization in the Slavic Languages», November 11–14, 2011. München-Berlin-Washington/D.C.: Verlag Otto Sagner. Band 55. P. 423–436. (In Russian).
9. *Iomdin L. L.* (2015). Constructions of Microsyntax formed with the Russian word raz // SLAVIA, časopis pro slovanskou filologii, ročník 84, 2015, sešit 3, s. 291–306.
10. *Iomdin, Leonid* (2016). Microsyntactic Phenomena as a Computational Linguistics Issue. // Grammar and Lexicon: Interactions and Interfaces. Proceedings of the Workshop. Osaka, Japan. 2016, pp. 8–18. (http://aclweb.org/anthology/W/W16/W16-38.pdf). ISBN 978-4-87974-706-8
11. *Iomdin L. L.* (2017a). What should we do about the constructions like kak byt? // // Kompjuternaja lingvistika I intellektual'nye tekhnologii. Po materialam ezhegodnoj Mezhdunarodnoj konferencii «Dialog» (2017). Moscow: Izdatelstvo RGGU. Issue 16 (23). Vol 1. p. 161–176. (In Russian.)

12. *Iomdin, Leonid* (2017b). Between the syntactic idiom and syntactic construction. Nontrivial cases of microsyntactic ambiguity. // SLAVIA, časopis pro slovanskou filologii, ročník 68, 2017, sešit 2–3, s. 230–243. (In Russian.)
13. *Jackendoff, Ray* (1997). Twisting the Night Away. // Language, Vol. 73, pp. 534–559.
14. *Marakasova A. A., Iomdin L. L.* (2016). Microsyntactic tagging in the SynTagRus corpus of Russian texts. [Mikrositaksicheskaja razmetka v korpuse russkikh tekstov SynTagRus] // Informatsionnye tekhnologii i sistemy 2016 (ITiS'2016). Sbornik trudov 40-j mezhdistsiplinarnoj shkoly-konferencii IPPI RAN. Repino, Saint Petersburg. p. 445–449. (In Russian).
15. *Mel'čuk, Igor* (2012). Phraseology in the language, in the dictionary, and in the computer. // Yearbook of Phraseology, Volume 3, Issue 1, pp. 31–56.
16. *Melikyan V. Yu.* Modern Russian Language. Syntax of the Indivisible Sentence. Teaching Manual. Rostov-on-Don. RGPU Publishers. 2004. — 288 p. (In Russian).
17. *Nikolaeva T. M.* (2008). (Nonparadigmatic linguistics. (The history of "straying particles") [Neparadigmaticheskaja lingvistika (Istorija "bluzhdajushikh chastic")]. Moscow: Jazyki slavjanskikh kultur. — 689 p. (In Russian).
18. *Shmelev D. N.* (1976). Syntactic divisibility of the utterance in Modern Russian. Moscow, Nauka, 152 p. (In Russian).
19. *Shvedova N. Ju.* (1958). On certain types of phraseologized constructions in the structure of Russian colloquial speech. // Voprosy jazykoznanija, № 2, p. 95–100. (In Russian).

EFFICIENCY OF TEXT READABILITY FEATURES IN RUSSIAN ACADEMIC TEXTS

Ivanov V. V. (nomemm@gmail.com)
Innopolis University, Innopolis, Russia

Solnyshkina M. I. (mesoln@yandex.ru),
Solovyev V. D. (maki.solovyev@mail.ru)
Kazan Federal University, Kazan, Russia

This paper addresses the problem of readability assessment for Russian texts and investigates the impact of 24 lexical, syntactic and frequency features. The research was conducted on Russian Readability Corpus containing two sub-corpora, two sets of 5–11 grade level textbooks on Social studies for native speakers of Russian. The sub-corpora were collected for research purposes, annotated and marked as BOG and NIK. The application of the Ridge regression has demonstrated the connection between readability and average sentence length, average number of coordinating chains, average number of sub-trees, frequency and lexical features. The results of the study have the potential to be applied in a wide variety of areas including primarily education, as well as webpage design, document management.

Key words: readability assessment, Russian Readability Corpus, average sentence length, average number of coordinating chains

ЭФФЕКТИВНОСТЬ ПРИЗНАКОВ ДЛЯ АНАЛИЗА СЛОЖНОСТИ АКАДЕМИЧЕСКИХ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

Иванов В. В. (nomemm@gmail.com)
Университет Иннополис, Иннополис, Россия

Солнышкина М. И. (mesoln@yandex.ru),
Соловьев В. Д. (maki.solovyev@mail.ru)
Казанский федеральный университет, Казань, Россия

В статье рассматривается проблема оценки удобочитаемости для российских текстов и исследуется влияние 24 лексических, синтаксических и частотных признаков. Исследование проводилось на российском корпусе (школьных) учебных текстов, содержащем два набора учебников уровня 5–11 класса по обществознанию для носителей русского языка. Две части корпуса, составляют учебники, написанные двумя разными авторами, были аннотированы и обозначены как BOG и NIK. Применение метода Ридж-регрессии продемонстрировало связь удобочитаемости со средней длиной предложения, средним числом координационных цепочек слов, средним количеством поддеревьев, частотой и лексикой. Результаты исследования могут быть применены в самых разных областях, включая прежде всего образование, а также дизайн веб-страниц, управление документами.

Ключевые слова: оценки удобочитаемости, Russian Readability Corpus, средняя длина предложения

1. Introduction

In the Russian Federation today, educators, parents and administrators are buzzing about Unified National Examinations, which are expected to mark a big shift to better practices of assessment. The latter is impossible if educators are not provided with a wide range of leveled reading materials to tailor all categories of students' learning programs. To achieve desired learning outcomes students and educators need available databases of leveled reading materials and textbooks to match various 'reader—text' profiles'. As for the textbook writers, they are expected to create books tailoring a wide range of abilities and goals but providing a minimal core syllabus for all categories of students (https://russian.rt.com/russia/article/434027-ministr-obrazovaniya-vasileva-intervyu). Special attention should also be paid to profiles of children with specific reading comprehension difficulties (https://alldef.ru/ru/articles/almanah-13/edinaja-koncepcija-specialnogo-federalnogo-gosudarstvennogo). Unfortunately, the existing textbooks which play the central role in teaching are traditionally criticized for being "nothing but collections of facts" (http://www.nlobooks.ru/node/2808) and for "complicated language" (https://www.znak.com/2014-04-08/pochemu_odin_iz_samyh_populyarnyh_uchebnikov_po_matematike_ne_proshel_gosudarstvennuyu_ekspertizu).

Realizing vital importance of reading for national progress, in 2003 Russia launched a sustainable "The National Program of Support and Development of Reading" which announces that "Modern Russia has approached a critical threshold in its neglect of reading on the national scale and at the moment we witness the beginning of the process of irreversible destruction of the nucleus of national culture" (http://www.library.ru/1/act/doc.php?o_sec=130&o_doc=1122). The program calls for evoking interest of younger generation in reading and turning Russians into "active readers". The Program also specifies the significance of "improving the quality and variety of readable literature in all areas of knowledge" and "establishing a system of selecting books for different categories of readers" (http://www.library.ru/1/act/doc.php?o_sec=130&o_doc=1122).

The research held in 2016–2017 showed that reading comprehension skills of Russian primary schoolchildren aged 9–10 top the list of international ranking, however, by the age of 15 Russian secondary schoolchildren gradually move to the middle of the ranking (http://docs.cntd.ru/document/436739637). All the above makes the problem of finding reading material of the right difficulty and assessing educational text readability relevant and even critical in realizing national goals.

As a part of a bigger research aimed at computing a readability formula for Russian texts, in this paper we address the following research question: what features in a linear regression model are informative for estimating readability of Russian academic texts.

2. Related work

Though studies on assessment of texts readability and readability formulas have a history of over a century [Chall, 1958], the Russian history of estimation of text readability is much shorter. Readability as a quantitative concept and a function of text variables was addressed for the first time as late as in the 1970s and 1980s (Lerner, 1974, Ushakov, 1980, Tomina, 1985, Tsetlin, 1980, Mackovskij 1976). By now Russian text analysts have five readability formulas at their disposal:

- Flesch Reading Ease Readability Formula
  $206.835 - (1.3 \times ASL) - (60.1 \times ASW)$

- Mikk [1970]: $0.01 \times x1 + 0.27 \times x2 + 0.54 \times x3$
- Mackovskij [1976]: $0.62 \times ASL + 0.123 \times X4 + 0.051$
- Tuldava (1975): $i \times lg(j)$,
- Oborneva (2006): $206.836 - (1.52 \times ASL) - (65.14 \times ASW)$

where:

- ASL, j = Average Sentence Length, the number of words divided by the number of sentences)
- S = the average number of sentences per 100 words.
- ASW, i = Average number of syllables per word, the number of syllables divided by the number of words),
- x1 = the length of sentences in the number of printed characters,
- x2 = the percentage of different unfamiliar words,
- x3 = the abstractness of the repetitive notions expressed by nouns,
- X4 = the percentage of more than 3-syllable words.

Though the threshold between short and long sentences or at least the number beyond which readability declines for all readers have never been adequately defined the average sentence and word length have always been viewed as good indicators of readability in the majority of readability formulas for Russian texts (see above).

Current studies on texts readability prove strong relationship between word frequency and text readability and provide concrete options for more effectively making

use of lexical frequency information in practice [Chen, X. B. & Meurers, D., 2016]. The results of Russian researchers' studies also show that text readability estimation should take into account the distribution of a range of lexical features in a text [see Mikk, 1970, Sharoff, 2008].

Extensive studies were also conducted on the impact of syntax on readability of Russian texts. As features influencing text readability, include the following: the number of participles, adverbial participles, the number of participial constructions, the number clauses in a complex and compound sentences. The researchers specifically emphasize the importance of different connectives such as conjunctions in compound sentences [Krioni, N. K., Nikin, A. D. & Filippova, A. V., 2008]. Far from being solved, the problem of readability correlation with text syntactic features still remains a challenging and highly relevant research area.

3. Feature analysis and model selection for text complexity prediction

In the model selection our main aim is to define an appropriate subset of features for a linear regression model. As described above such models are well-known and are based on two or three parameters (such as average sentence length or average syllables per word). To the best of our knowledge there was no investigation of a wider set of features for prediction of text complexity in Russian. Given the small number of texts in the corpus, we are focus on those linear models, that will not overfit.

3.1. Description of features

In this paper we explored an extended feature set for text complexity modeling. The first part of the feature set contains features based on length and frequency. This part includes 'average words per sentence', 'average syllables per word' and 'frequency of content words' (FREQ). The FREQ feature is calculated using the Russian frequency dictionary. We count frequencies for each word in an input text.The second part of feature set includes features calculated from part-of-speech tags. In fact, these features represent number nouns, adjectives, verbs, pronouns and negations occur in a text. The POS-tags were derived using TreeTagger (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/). The third part includes syntactic features derived with ETAP-3 system.

PART1: Features based on length and frequency:

- FREQ is a cumulative frequency of content words,
- ASL is an average number of words per sentence,
- ASW is an average number of syllables per word.

PART2: Features based on POS tags:

- NOUNS is a number of nouns per sentence,
- VERBS is a number of verbs per sentence,

- ADJ is a number of adjectives per sentence,
- PRONOUNS is a number of pronouns per sentence,
- PERONAL PRONOUNS is a number of personal pronouns per sentence,
- NEG is a number of negations per sentence.

PART3: Features based of syntactic dependencies:

- AVERAGE_PATH is the quotient of the number of nodes and the number of leaves in a sentence.
- AVERAGE_SOCHIN_LENGTH is the average length of coordinating constructions
- DEEPRICH_RATE is the average number of verbal participles.
- DEEPRICH_V is the average span of a verbal adverb phrase.
- LEAVES_NUMBER is the average number of 'leaves' (terminal nodes, i.e., words that are not anyone's "hosts") in a sentence.
- LONGEST_PATH is the average length of the longest branch.
- NOUNS_DEP is the average number of modifiers in a nominal group; coordinating and explanatory links are ignored.
- PODCHIN_NUMBER is the ratio of sentences in which there is at least one subordinate conjunctions or relational links.
- PODCHIN_RATE is the average number of subordinate links.
- PRICH_RATE is the average number of participial construction; participial constructions are defined as a participle that has at least one dependent.
- PRICH_V is the average span of a participial construction is the quotient of the number of nodes that depend on the participle.
- SENTSOCH_NUMBER is the average number of compound sentences.
- SOCHIN_NUMBER is defined as the average number of coordinating chains.
- PATH_NUMBER is defined as the average number of sub-trees (in a sentence).
- VERBS_DEP is defined as the average number of finite dependent verbs and is calculated as the sum of nodes directly dependent on the finite verb divided by the number of finite verbs; coordinating and explanatory links were ignored.

3.2. Description of the corpus

The first major statistical issue in building a corpus of texts, as Biber (1990) puts it, "concerns the sampling of texts: how linguistic features are distributed across texts and across registers, and how many texts must be collected for the total corpus and for each register to represent those distributions?" Having compared the internal variations of the two texts in the corpus, Biber (1990) concludes that text samples of 1000 words are representative for the text categories under study. He also proved that the 20–80 samples of texts are enough for correlation-based analysis [Biber 1990].

Two collections of texts were assembled for the research. The first collection of 7 texts derived as a result of OCR and postprocessing of textbooks on Social Studies by L. N. Bogolubov. We mark this collection as "BOG". Textbooks cover range of 6–11 Grade Levels. The second collection of 7 texts from textbooks on Social Studies by A. F. Nikitin marked "NIK" aimed at 5–11 Grade Levels. Further we refer to the

two collections collectively as a Russian Readability Corpus (RRC). Both sets of textbooks are from the "Federal List of Textbooks Recommended by the Ministry of Education and Science of the Russian Federation to Use in Secondary and High Schools". To ensure reproducibility of results, we uploaded the corpus on a website (http://kpfu.ru/slozhnost-tekstov-304364.html). Note, however, that the published texts contain shuffled order of sentences. This shuffling, indeed, does not affect the values of features, because they do not depend on sentence order. Table 1 provide numerical description of the RRC.

Table 1. Numerical Data on RRC

Grade level	Tokens		Sentences		ASL		ASW	
	BOG	NIK	BOG	NIK	BOG	NIK	BOG	NIK
5-th	—	17,221	—	1,499	—	11.49	—	2.35
6-th	16,467	16,475	1,273	1,197	12.94	13.76	2.56	2.71
7-th	23,069	22,924	1,671	1,675	13.81	13.69	2.84	2.70
8-th	49,796	40,053	3,181	2,889	15.65	13.86	2.96	2.88
9-th	42,305	43,404	2,584	2,792	16.37	15.55	3.04	3.00
10-th	75,182	39,183	4,468	2,468	16.83	15.88	3.07	3.12
10-th*	98,034	—	5,798	—	16.91	—	3.05	—
11-th	—	38,869	—	2,270	—	17.12	—	3.11
11-th*	100,800	—	6,004	—	16.79	—	3.19	—

Comment. Star sign (*) denotes advanced versions of books for the corresponding grade; sign '—' denotes absence of a textbook for the corresponding grade.

3.3. Analysis of features

3.3.1. Correlation between features

We provide the results of correlation analysis in the Table 2. In general, some syntactic features are similar to others and correlate with the target variable (readability, measured as a grade level). However, it is evident that all the syntactic features have lower correlation coefficient with the target feature ('Grade Level'), than the two 'classical' lexical features (ASL and ASW) do.

Table 2. Correlation between features and target feature, grade level

	Feature name	Correlation coefficient		Feature name	Correlation coefficient
1	ASL	0.94	6	AVERAGE_SOCHIN_LENGTH	0.87
2	ASW	0.94	7	PATH_NUMBER	0.87
3	SOCHIN_NUMBER	0.93	8	LONGEST_PATH	0.84
4	PRICH_RATE	0.91	9	FREQ	0.84
5	NOUNS_DEP	0.88			

	Feature name	Correlation coefficient
10	LEAVES_NUMBER	0.84
11	AVERAGE_PATH	0.84
12	ADJ	0.82
13	NOUNS	0.82
14	VERBS	0.74
15	NEGATIONS	0.70
16	PRONOUNS	0.70
17	PODCHIN_RATE	0.64
18	PODCHIN_NUMBER	0.62
19	DEEPRICH_V	0.52
20	PERS_PRONOUNS	0.47
21	DEEPRICH_RATE	0.44
22	VERBS_DEP	0.43
23	PRICH_V	0.33
24	SENTSOCH_NUMBER	0.03

3.3.2. Significance of features

We tested significance of a linear regression model in the following setting. We applied the F-test for linear regression to evaluate whether any of the independent variables in a multiple linear regression model are significant. The results of the F-test are presented in the table below. P-values are denoted with '**' and '*' signs.

Table 3. Results of F-test for significance of attributes of a linear regression model (** corresponds to p-values < 0.01; * corresponds to p-values < 0.05)

	Feature name	F-score
1	ASL	95.58**
2	ASW	91.93**
3	SOCHIN_NUMBER	71.23**
4	PRICH_RATE	56.20**
5	NOUNS_DEP	42.17**
6	AVERAGE_SOCHIN_LENGTH	38.91**
7	PATH_NUMBER	35.69**
8	LONGEST_PATH	29.45**
9	FREQ	29.32**
10	LEAVES_NUMBER	29.01**
11	AVERAGE_PATH	28.60**
12	ADJ	25.33**
13	VERBS	24.49**
14	NOUNS	19.17**
15	NEGATIONS	14.11**
16	PERS_PRONOUNS	11.00**
17	PODCHIN_RATE	8.35*
18	PODCHIN_NUMBER	7.41*
19	DEEPRICH_V	4.49
20	DEEPRICH_RATE	2.86
21	VERBS_DEP	2.76
22	PRICH_V	1.42
23	PRONOUNS	0.22
24	SENTSOCH_NUMBER	0.01

It was expected that, that the most significant attributes include well-known features on length of sentences and words (ASL, ASW), syntactic features (such as SOCHIN_NUMBER, PRICH_RATE, etc.) and lexical attributes (ADJ, VERBS, etc.).On the other hand, insignificant features include SENTSOCH_NUMBER, PRICH_V, etc. which corresponds to correlation analysis. We use results of this evaluation for filtering insignificant features. Therefore, based on p-value (<0.01), for further analysis we keep only first 16 features from the Table 3. It is clear that 16 features are too many to build a robust linear regression given the number of texts in our corpus.

In the next step we make use a technique for feature selection: Ridge regression [Wessel N. van Wieringen, 2018] to find a subset of most relevant features for a prediction model. An alternative is just a brute-force search for the best subset of features. A drawback of thebrute-force approach is clear: given the number of texts in the corpus a model with many featurescan easily overfit the data even if we split the dataset into a train and test sets.

3.3.3. Feature selection with Ridge regression

Ridge regression is an approach that represents regularization technique with constrain (L2-norm) on the feature weights in a linear model. The approach can be used to rank features with respect to their magnitude (their influence on the target variable). We use the ranked list of features to select reasonable subset of features for linear regression model of text complexity.

Table 4. Ridge regression results in feature selection

	Feature	Absolute value of Coefficient in Ridge Regression		Feature	Absolute value of Coefficient in Ridge Regression
1	**ASL**	0.506	9	NOUNS_DEP	0.071
2	ASW	0.125	10	FREQ	0.034
3	SOCHIN_NUMBER	0.119	11	NEGATIONS	0.010
4	PRICH_RATE	0.106	12	AVERAGE_PATH	0.007
5	LONGEST_PATH	0.089	13	PERS_PRONOUNS	0.003
6	PATH_NUMBER	0.079	14	VERBS	0.001
7	LEAVES_NUMBER	0.075	15	ADJ	0.001
8	AVERAGE_ SOCHIN_LEN	0.071	16	NOUNS	0.000

4. Discussion of results and conclusion

With the view of increasing amount of available academic texts, broadening varieties of alternative training options and personalized training, the problem of selecting appropriate teaching materials is becoming urgent. Textbooks of almost the same content may differ in the degree of complexity (readability) of presentation. To the best of our knowledge, there have been no extensive multi-feature studies of readability of Russian texts. The authors of the paper offer an innovative 24-feature analysis of Russian texts readability embracing “classical” frequency features, part-of-speech, and syntactic features. For our research we create dataset which are uploaded on KFU website and are available for potential verification and validation of the research outcomes.

The results derived in this paper support the following points. First, average sentence length is the most important feature for text complexity prediction. Second, there are several highly important syntactic features such as the average number of coordinating chains, average number of sub-trees, as well as frequency and lexical

features that can improve prediction. Third, surprisingly, average syllables per word may not be a very important feature (in presence of other features), even though it correlates with target variable.

The results obtained in this article are far from being final, since they are received on a relatively small corpus of homogeneous texts. Readability of different types of texts is to be estimated with different formulas. Rather, this article offers a methodology for this type of research. We intend to further apply the proposed approach to texts of other subject areas and genres. It is also proposed to further expand the set of text features to be studied, including semantic and discursive features. The research available suggest that lexical features of reading texts such as word frequency, word identification ability, mean noun frequency level as well as lexical diversity and type-token ratio (TTR) are factors that influence reading comprehension, it is this fact that makes them reliable metrics in assessing text complexity [see Solovyev 2018]. Based on the hypothesis that the average word frequency across the textbooks is to have consistent progression, we plan to conduct a cross-sectional (grade) study of textbooks for different age groups with regard to lexical density, TTR and lexical diversity.

Though selecting appropriate reading text for students of different grades is of crucial importance, we do not narrow our studies to educational needs only. The research shines a light on issues worthy of discussion with regard to texts used in mass media, healthcare, document management, etc. Contributing this article we hope to attract attention of scholars working in related areas so that we could combine our efforts and change the opportunities for thousands of struggling readers. National discussions are needed to ensure that writers (textbook authors, speech and news writers, journalists, etc.) can make informed decisions about the difficulty level of the texts they generate.

Acknowledgements

This research was financially supported by the Russian Science Foundation, grant № 18-18-00436, the Russian Government Program of Competitive Growth of Kazan Federal University, and the subsidy for the state assignment in the sphere of scientific activity, grant agreement № 34.5517.2017/6.7. The Russian Academic Corpus (section 3.2 in the paper) was created without supporting by the Russian Science Foundation. We would like to convey our sincere gratitude to Dr. Lana Timoshenko and Dr. Ivan Rygaev for their valuable assistance with syntactic annotation of the corpus.

References

1. *Biber, D.* (1990). Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation. Literary and Linguistic Computing, Literary and Linguistic Computing, 5: 257–69.
2. *Chall, J. S.* (1958). Readability: An appraisal of research and application. Bureau of Educational Research Monographs, No. 34. Columbus, Ohio State Univ. Press.
3. *Chen, X. B., Meurers, D.* (2016). Characterizing Text Difficulty with Word Frequencies. In Proceedings of The 11th Workshop on Innovative Use of NLP for

Building Educational Applications (pp. 84–94). San Diego, CA. Association for Computational Linguistics.
4. *Choldin, M. T.* (1979). Rubakin, Nikolai Aleksandrovic. In A. Kent, H. Lancour, J. E. Daily (Ed.) Encyclopedia of library and information science. (pp. 178–179). Basel: CRC Press.
5. *Krioni N. K., Nikin A. D., Filippova A. V.* (2008). Avtomatizirovannaya sistema analiza parametrov slozhnosti uchebnogo teksta In Tekhnologiya i organizatsiya obucheniya : nauch. izdanie. — Ufa: UGATU, P. 155–161.
6. *Lerner, I. Ya.* (1974) Kriterii slozhnosti nekotorykh elementov uchebnika: Problemyshkol'nogouchebnika [The criteria for the complexity of some elements of the textbook: Problems of a school textbook]. Is. 1. Moscow: Prosveshchenie.
7. *Mackovskij, M. S.* (1976). Problemy chitabel'nosti pechatnogo materiala. Smyslovoe vospriyatie rechevogo soobshcheniya v usloviyah massovoj kommunikacii [Problems of readability of printed material. Semantic perception of speech messages in conditions of mass communication]. Moskow: Nauka.
8. *Mikk, Y. A.* (1970) O factorakh ponyatnosti uchebnogo teksta [On factors of comprehensibility of educational texts]. Diss. ... cand. ped. sciences. Tartu.
9. *Mutt, O.* (1984). Aktual'nye voprosy otbora uchebnogo materiala dlya vuzovskogo kursa inostrannogo yazyka [Actual questions of selection of the educational material for the university course of a foreign language]. Tartu: The Tartu State University
10. *Oborneva, I. V.* (2006) Avtomatizirovannaya otsenka slozhnostiuchebnykh tekstov na osnove statisticheskikh parametrov [Automated estimation of complexity of educational texts on the basis of statistical parameters]. Pedagogy Cand. Diss. Moscow.
11. *Sharoff, S., Kurella, S. and Hartley, A.* (2008). Seeking needles in the web's haystack: Finding texts suitable for language learners. In Proceedings of the 8th Teaching and Language Corpora Conference (TaLC-8).
12. *Solovyev, V., Ivanov, V., Solnyshkina M.* (2018) Assessment of reading difficulty levels in Russian academic texts: Approaches and Metrics. Journal of Intelligent & Fuzzy Systems, (Preprint):1–10, 2018.
13. *Tomina, Yu. A.* (1985) Ob"ektivnaya otsenka yazykovoy trudnosti tekstov (opisanie, povestvovanie, rassuzhdenie, dokazatel'stvo) [An objective assessment of language difficulties of texts (description, narration, reasoning, proof)]. Abstract of Pedagogy Cand. Diss. Moscow.
14. *Tsetlin, B. C.* (1980) Didakticheskie trebovaniya k kriteriyam slozhnosti uchebnogo materiala [Didactic requirements to the complexity criteria of educational material]. Novye issledovaniya v pedagogicheskikh naukakh. 1 (35). pp. 30–33.
15. *Tuldava, Yu. A.* (1975) Ob izmerenii trudnosti tekstov [On measuring the complexity of the text]. Uchenye zapiski Tartuskogo universiteta. Trudy po metodike prepodavaniya inostrannykh yazykov. 345. pp. 102–120.
16. *Ushakov, K. M.* (1980) O kriteriyakh slozhnosti uchebnogo materiala shkol'nykh predmetov [On the criteria of complexity of teaching material of school subjects]. Novye issledovaniya v pedagogicheskikh naukakh. 2 (36). pp. 33–35.
17. *Wessel N. van Wieringen* (2018) Lecture notes on ridge regression. arXiv:1509.09169v2 [stat.ME] 6 Jan 2018.

CORPUS-BASED INVESTIGATION OF QUOTATION IN RUSSIAN SIGN LANGUAGE[1]

Khristoforova E. A. (evkhristoforova@gmail.com),
Kimmelman V. I. (vadim.kimmelman@gmail.com)

Russian State University for the Humanities, Moscow, Russia

This paper presents corpus-based research of quotation constructions in Russian Sign Language (RSL). Quotation constructions have been observed from different perspective in different signed and spoken languages [Brendel, Meibauer, Steinbach 2011]; [Litvinenko et al. 2009]. Based on the corpus of spontaneous narratives recorded from RSL signers [Burkova 2015], we conducted a quantitative analysis of these constructions. We analyzed constituents of quotation construction, such as the source (author of utterance) indication, the introducing matrix predicate, and the quote. Our investigation of non-manual markers in the corpus revealed that non-manual marking of quotation is optional for RSL quotations. We distinguished direct and indirect quotations in our data based on the reference of indexical elements, the use of subordinating conjunction, and the imperative mood. We found that in RSL non-manuals do not mark the direct/indirect type of quotation. Our data show that RSL signers tend to use direct quotation much more frequently than indirect quotation. In addition, we compared our findings with the data on quotation constructions in some other sign languages and with the studies of quotation in natural discourse of spoken languages. This comparison showed that RSL quotations share core properties with quotations in spoken and signed languages [Litvinenko et al. 2009].

Key words: quotation, sign languages, RSL, corpus-based research, non-manual markers

КОРПУСНОЕ ИССЛЕДОВАНИЕ ЦИТАЦИОННЫХ КОНСТРУКЦИЙ В РУССКОМ ЖЕСТОВОМ ЯЗЫКЕ

Христофорова Е. А. (evkhristoforova@gmail.com),
Киммельман В. И. (vadim.kimmelman@gmail.com)

Российский государственный гуманитарный университет, Москва, Россия

[1] This work was supported by the RSF grant 17-18-01184.

В данной статье представлено корпусное исследование цитационных конструкций в русском жестовом языке (РЖЯ). Цитационные конструкции были изучены с разных точек зрения на материале как жестовых, так и звуковых языков [Brendel, Meibauer, Steinbach 2011; Litvenenko et al. 2009]. Для настоящего исследования мы использовали корпус спонтанных нарративов, записанных от носителей РЖЯ [Burkova 2015]. Анализ корпуса позволил количественно описать такие составляющие цитационных конструкций, как указание на автора высказывания, вводящую предикацию и собственно саму цитацию. В процессе анализа немануальных маркеров, представленных в корпусе, было обнаружено, что немануальное маркирование цитаций не является обязательным в РЖЯ. В нашем корпусе мы разделяли прямую и косвенную цитацию, основываясь на следующих критериях: сдвиг референции индексикалов, наличие подчинительного союза и показателей императива. Мы обнаружили, что различие между прямым и косвенным типом цитации не маркируется немануально. Мы отметили, что носители РЖЯ используют прямую цитацию значительно чаще, чем косвенную. Сравнив наши результаты с данными исследований цитационных конструкций в других жестовых языках и в естественном дискурсе звуковых языков, мы пришли к выводу, что цитационные конструкции РЖЯ имеют много общего с цитацией в звучащих и жестовых языках [Litvinenko et al.2009].

Ключевые слова: цитация, жестовые языки, РЖЯ, корпусное исследование, немануальные маркеры

1. Introduction

Quotation constructions, that is, the means of conveying other's words and thoughts have been the topic of numerous studies [Brendel, Meibauer, Steinbach 2011]. Typically, direct and indirect quotation are distinguished. Direct quotation is almost verbatim representation of an utterance, while indirect quotation is a report from narrator's perspective. In writing, one may observe the difference in punctuation between these two types. In natural discourse, different intonation can be used for direct or indirect quotation. In addition, these two types of quotation differ in the reference of indexicals—elements whose reference depends on the context, such as personal pronouns and time adverbials. In direct speech, indexicals must be interpreted within the context of the quoted situation, while in indirect quotation they are interpreted within the overall context of the narration. There are also structural differences between direct and indirect quotation. Indirect quotation in most languages tends to be expressed by an embedded clause. Direct quotation, on the contrary, is syntactically independent. Notwithstanding all the differences between the types of quotation, it is not always easy to distinguish direct and indirect quotation in natural discourse [Litvinenko et al. 2009].

Sign languages also have quotation constructions. Investigating quotation in sign languages, researchers observed a phenomenon called "role shift" [Herrmann & Steinbach 2012; Quer 2011; Schlenker 2017]. Signers tend to shift into the role of a character using the range of non-manual markers. Among these non-manual markers are leans or turns of the body and head, change of eye gaze direction, and

different emotional facial expressions. Performing role shift signers not only sign from a character's point of view conveying his words, emotions, or thought, but they can also act from character's perspective presenting his actions. This is different from the quotation in spoken languages [Liddell & Metzger 1998].

Role shift has been explored from different perspectives. Some researchers concluded that the use of role shift does not always clearly indicate direct quotation, but shows some properties of indirect speech. Much research has been done to identify non-manual markers, which accompany quotation in sign languages [Herrmann & Steinbach 2012 i.a.].

This paper describes quotation in Russian Sign Language (RSL). We identified quotation constructions in a corpus of spontaneous narratives. We described the constituents of quotation constructions, the most frequent non-manual markers, and the differences between direct and indirect quotation. We also compared our findings with quotation in other sign languages and in natural discourse of spoken languages.

2. Methodology

We have chosen to investigate quotation in RSL based on corpus data. Corpus-based research methods have only recently been applied to sign languages (see [Lucas, Bayley and Valli 2001] for one of the first studies). Corpus research gives a possibility for quantitative studies of natural language, and sign languages are not the exception. Modern computer-based devices, such as ELAN or SignStream, provide sign language linguists with tools for multilayered annotation [Safar & Glauert 2012]. Considering multi-modal nature of sign languages, these tools are essential for full description and profound investigation of sign languages. There are also structured search engines that allow users to study overlap between distinct layers of description. A common obstacle that sign language linguists are bound to deal with is the lack of automatic annotation devices. Tagging, parsing and transcription are all completely manual. That makes annotation process complicated, time-consuming and highly annotator specific. In addition, manual factor leads to the low productivity of the annotation process, which may excuse relatively small sizes of most corpora [Safar & Glauert 2012].

On-line RSL corpus was created by a team of sign language researchers lead by [Burkova 2015]. Currently, this corpus is the only RSL corpus available for public use. It contains around 200 video recordings of different text types (picture-based storytelling, interviews, spontaneous narrations, elicitations, etc.). Most recordings are annotated in ELAN software[2] using four layers: glosses for the right hand signs, glosses for the left hand signs, overall translation of the clause and comments (**Figure 1**).

[2] https://tla.mpi.nl/tools/tla-tools/elan/

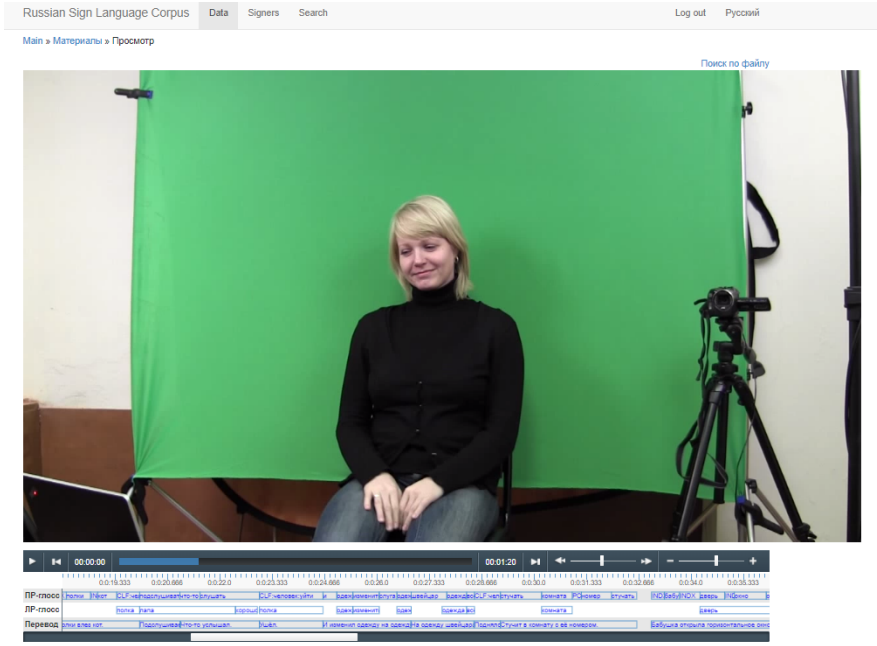


Fig. 1. Screenshot of the website

RSL is spread across the large territory of Russian Federation, which necessarily leads to some dialectal variation (Schembri & Johnston 2012). On-line RSL corpus mostly represents Moscow and Novosibirsk variants of RSL. These two sections of data must be carefully analyzed in order to sort out the influence of dialectic variation, but this goes beyond the purposes of thisstudy. For this study, we have chosen spontaneous narrations recorded in Moscow. This part of the corpus contains videos recorded from eleven RSL signers.

The chosen section of corpus contains almost 8,000 signs, forming nearly 1,200 clauses. Although this amount of data is relatively small, it contains 341 quotations. We would like to highlight that we consider not only reporting the speech of others but also reporting thoughts and attitudes. Conveying these meanings, speakers and signers also use direct or indirect quotation. Using ELAN, we annotated all the videos using the following tiers:

- Type of quotation (speech/thought/attitude);
- Matrix verb, introducing quotation: the verb of speech (or thought/attitude) if used;
- Source of quotation: whether the author of quotation is lexically introduced or not. If there is no lexical sign for the source but it matches the subject of the preceding action enacted by the signer, the source may be easily reconstructed from the context and its lexical representation would be excessive. We indicated such cases as “constructed action”;

- Author of quotation: whether the signer himself is the author of quotation (s/he uttered it in the past) or some other person uttered it;
- Non-manual markers: eye gaze direction (distinct from the one towards the interlocutor), head turns, and body leans and turns;
- Indexicals: indexical elements (deictic elements, time indications, agreeing verbs personal pronouns, possessive markers) in the quotation and their reference (shifted or not);
- Subordination features: elements moved from the quotation to the matrix clause;
- Markers of direct or indirect speech: imperative markers or conjunctions.

The non-manual markers we annotated (eye gaze, head turns, body leans and turns) are the ones most commonly associated with quotation in sign languages [Herrmann & Steinbach 2012]. Another common marker is emotional facial expressions attributable to the author of the quote. However, various researchers (ibid. a.o.) have demonstrated that facial expressions cannot be analyzed as markers of quotation as they are a part of the quoted utterance; they also clearly occur without quotation. We thus left them out of our analysis.

3. Properties of quotation in RSL

For RSL quotation, we have revealed the same basic constituents as previously identified for quotation constructions in other languages: the indication of the source (author) of quotation; introducing matrix predicate; the quotation itself (1).

(1) I_{source} $\text{SAY}_{\text{matrix predicate}}$ $[\text{I YES-YES VIA MOSCOW GO\_BY\_TRAIN}]_{\text{quotation}}$
I say: "Yes, I go via Moscow by train".

Source indication and introducing matrix predicates are optional. Source is indicated in 160 (47%) cases, while in 118 (35%) cases it is not mentioned at all. Authors of other 52 (15%) quotations matched subjects of previously described actions, so that the source of these quotations was not expressed lexically, but was easily retrieved from the context. According to [Mathis & Yule 1994], an optional source indication is also common for spoken languages.

Introducing matrix predicates share the same property of optionality. 218 (64%) quotations are not introduced by any predicate. We have also found that in 27 (8%) cases the quotation is proceeded by a sign called "palms up" instead of matrix predicate (2). We assumed that this sign might also be analyzed as a quotation marker. In RSL "palms up" is considered to be a multifunctional sign without any specific lexical meaning. Although we suggest that it is used to introduce quotations in our data, we still cannot exclude that this sign may bear other functions. Interestingly, according to our data, "palms up" tend to introduce quotations whose author is the signer himself but in past. Optional indication of matrix predicates of quotation is also among the properties of quotation in spoken languages [Litvinenko et al. 2009].

(2) I PALMS_UP YES YES OKAY IMPERATIVE
I say: "Yes, okay, let's go!"

Figure 2. "Palms up" sign

As for 123 (36%) cases of matrix predicates of quotation, the most frequent are THINK, TELL, SAY, ASK and CALL. In addition, we have identified a class of verbs that introduce quotation quiet frequently but do not belong to verbs of speech or thoughts such as CALL: syntactically, it introduces quotation, but in fact, it describes the action preceding quotation (3).

(3) DAVYDENKO(proper noun) CALL-1 NUMBER BOX WHAT BOX
Davydenko calls me: "What is the number of box?"

Non-manuals are another typical marker of quotation. As was already mentioned in section 2, in this paper we analyze eye gaze, head turns, and body movements and do not take into account facial expressions (4). Non-manual marking in RSL also turned out to be non-obligatory. 195 (57%) quotations are marked by the change of eye gaze direction, 175 (51%) are marked by body movements, and 287 (84%) by head turns. Head turns thus seem to be the most reliable marker, but see below. We have found 95 (26%) quotations without gaze or body turns, and 16 (5%) that simply have no non-manual marking.

(4) <u>eye gaze, head turn, body lean</u>
CALL-3 INTERESTING THIS
I call her: "This is interesting!"

Figure 3. Non-manual markers of quotation

In our data, non-manuals in RSL, if they are present, do not always accompany only quotation itself as typically observed for other sign languages. As shown in Table 1, only in 47 (14%) cases for eye gaze, in 58 (17%) cases for body movements, and in 62 (18%) for head turns these non-manuals mark the whole quotation and nothing else. Our data shows that non-manuals in RSL can also accompany other constituents of quotation constrictions. In fact, these results challenge our assumption that eye gaze and body movement mark quotations in all the cases.

Table 1. Non-manual marking of different constituents of quotation constructions

	Eye gaze	Body movement	Head turns
Part of quotation	52	63	62
Whole quotation	47	58	62
Whole quotation + matrix predicate	38	22	46
Whole quotation + matrix predicate + source indication	58	32	117

Note that body leans are relatively more often used to mark the whole quotation than eye gaze (χ^2=12.9, p = 0.005) which might indicate that it is a better marker overall (although also highly optional). Head turns are also different significantly from both gaze and body movement because they are very often used even on the source. This shows that they are not really good markers of the quote itself even though they are very frequent.

Considering all our findings discussed above, we tried to come up with a hypothesis explaining why these markers are present in some cases and absent in others. For example, we checked whether non-manual marking is connected with the report of thoughts instead of the report of speech (Table 2). We found that body movement is used significantly more often for reported thought than speech ($\chi^2 = 9.6$, p = 0.02). Importantly, body movements are used with both types. The differences for eye gaze and head turns were not significant. We also proposed that non-manual markers are less frequent when the signer himself is the author of quotation. However, this hypothesis was not supported by our data either: no differences are statistically significant (Table 3).

Table 2. Non-manual marking of different types of quotation

	Eye gaze	Body movement	Head turns	Total
Reported speech	155 (56%)	131 (47%)	245 (88%)	277
Reported thoughts/ attitudes	40 (62%)	44 (68%)	52 (81%)	64

Table 3. Non-manual markers and source of quotation

	Eye gaze	Body movement	Head turns	Total
Author of quotations is the signer	107 (61%)	94 (55%)	141 (82%)	171
Author of quotation is not the signer	99 (58%)	86 (51%)	147 (86%)	170

We further hypothesized that non-manual markers would be used more frequently when not a single utterance but a whole dialogue is quoted. We found nineteen cases of reported dialogues in the corpus. It turned out that that non-manual marking of quotations within the dialogues is similar to the overall pattern, so this hypothesis was not confirmed.

Finally, we hypothesized that non-manual marking correlates with the type of quotation (direct or indirect). Before we test this prediction, it is necessary to present our findings concerning direct and indirect speech in RSL.

4. Direct and indirect quotation

As was already mentioned in section 1, we can identify direct or indirect quotation basing on the following basic criteria:

- Reference of indexicals (shifted/non-shifted)
- Syntactic status of quotation as a sentence (embedded clause/independent clause).
- Possibility of imperative quotes (possible/impossible)

First, we analyze indexical elements in our data. Most frequent of them are personal pronouns (I, YOU, HE/SHE); possessive markers; agreeing verbs; time and place adverbials like HERE, and NOW; tense markers for past and future.

It turned out that 196 (57%) quotations do not have indexical elements. Another obstacle is that, in the case of the signer quoting his own speech in the past, the first-person pronoun can refer to the signer in the context of signed quotation (as the author of quotation) and to the signer in the context of general narration within which the quotation is reported. In such cases, it is impossible to define the reference of the pronoun as shifted or not.

Considering indexical element whose reference can be identified without complications, we found that 95% (86/91) of them have shifted reference. It implies that

for the quotations with the use of indexical elements direct quotation type is much more frequent in RSL.

As far as we have at least 86 examples of direct quotations and 8 examples of indirect quotations, we can proceed to the investigation of the connection between non-manual marking and the type of quotation. We hypothesized that non-manuals might only mark direct quotations, as direct speech is supposed to be more emotional (as it is in spoken languages). However, our data shows that the use of non-manuals is not defined by the type of quotation. Among 86 direct quotations, we have found 31 quotations without eye gaze, 30 quotations without body movements, 8 without head turns, and 3 without any non-manual marking. Also recall that in many examples, including those with shifted indexicals, the non-manuals are not aligned with the quote. Thus, non-manual markers are not obligatory for direct quotations.

Apart from the reference of indexicals, we also analysed features of subordination. The most obvious marker of quotation represented by an embedded clause is a subordinating conjunction THAT. This conjunction is used after matrix predicate in order to introduce embedded quotation. We assume that if THAT is used to introduce particular quotation, this quotation should be considered as indirect.

Constructions with the use of subordinating conjunction THAT are quite rare in our data: only seven cases were found (5). Most likely such constructions are borrowed from Russian. It is also important to note that among seven quotations with subordinating conjunction some are accompanied by non-manual markers. This is another proof that non-manuals do not only mark direct quotations.

(5) EXPLAIN THAT DEFINITELY THERE PST GO PST
We explain that we definitely came.

Our last criterion to identify the type of quotation is the use of imperatives quotes. According to [Kuno 1988], imperatives within quotation are a clear attribute of direct speech. Although not all researches agree with this statement, we decided to examine all the cases of imperatives within quotations in our data in order to investigate whether the use of imperatives correlate with the type of quotation. We have found 15 cases of the use of imperative manual marker within quotations. None of these quotations are introduced by subordinating conjunction and none of them contain indexical elements with non-shifted reference. Consequently, we consider them direct quotations, which implies that imperatives can be considered as a direct speech indication in RSL.

Summing up this section, we can give the following description of direct and indirect quotation constructions in RSL: direct quotation is the most frequent pattern to convey somebody's words/thoughts/attitude in RSL. It is explicitly indicated by the shifted reference of indexical elements, by the lack of subordinating conjunction THAT, or by imperatives within quotation. Our assumption about non-manual marking of direct or indirect quotations has not been confirmed: according to our data non-manual marking is optional and not fully aligned with quoteseven in direct speech.

5. Summary

Our aim was to describe quotation constructions in RSL based on corpus materials. Within on-line RSL corpus we have chosen spontaneous narrations, recorded in Moscow. In this part of data, we found 341 quotations. Using ELAN, we annotated each quotation by tiers listed in section 2, which allowed us to investigate the properties of quotation constructions in RSL from different perspectives.

Quotation constructions in RSL consist of the same elements as quotations described in other natural languages. These elements are source (author) indication, introducing matrix predicate, and the quotation itself. Source indication and introducing matrix predicate are optional. Although at least one of these elements accompanies the majority of quotations in our data, it is possible to find quotation construction consisting only of quotation: 12% (40) of quotation constructions in our data do not have either source indication or introducing matrix predicate.

We also investigated non-manual marking of quotations in RSL. Although many sign language researchers state that non-manual marking of quotations is obligatory or at least highly frequent [Herrmann & Steinbach 2012], our data shows that for RSL quotations non-manual markers are optional. The most common marker is the head turns but it usually marks the source and (a part of the) quote, not just the quote.

Following generally accepted criteria of direct/indirect speech (reference of indexical elements, syntactic structure, and possibility of imperatives), we found that for more than a half of quotations in our data we cannot define the type of quotation. Among those quotations that can be identified as direct or indirect, the majority of quotations (95%) are direct. RSL shares this property with spoken languages, in which direct quotations are also more frequent [Litvinenko et al. 2009]. This fact may be also explained by the relatively young age of RSL. It is possible that syntactic structure of indirect quotation in RSL may be in its early stage of grammaticalization [Pfau et al. 2016], but such processes are still understudied.

We also hypothesized that non-manual markers may indicate the type of quotation (that is, that only direct speech will have them). However, our data contradicted this hypothesis as described in detail in section 4. Difficulties with identifying the type of quotation are not unique for RSL: natural discourse of spoken languages has the same property [Litvinenko et al. 2009].

References

1. *Brendel E., Meibauer J., Steinbach M.* (2011), Exploring the meaning of quotation, Understanding Quotation, Berlin: De Gruyter Mouton, pp. 1–34.
2. *Burkova, S.* (2015), Russian Sign Language Corpus: http://rsl.nstu.ru/.
3. *Herrmann A., Steinbach M.* (2012), Quotation in sign languages: A visible context shift, Quotatives. Cross-linguistic and Cross-disciplinary Perspectives, Amsterdam: John Benjamins, pp. 203–230.
4. *Litvinenko, A., Korotaev N., Kibrik A., Podlesskaya V.* (2009), Constructions with quotation or "reported speech" [Konstruktsii s tsitatsiey, ili "chuzhoi rechyu"], Night Dream Stories: A Corpus Study of Spoken Russian Discourse [Rasskazi

o snovedeniah: Korpusnoye issledovanie ustnogo russkogo diskursa], Moscow: Language of Slavic Cultures, pp. 288–307
5. *Kuno S.* (1988), Blended quasi-direct discourse in Japanese, Papers from the Second International Workshop on Japanese Syntax, Stanford: CSLI, pp. 75–102.
6. *Liddell S. K., Metzger M.* (1998), Gesture in sign language discourse, Journal of Pragmatics, Vol. 30, № 6, pp. 657–697.
7. *Lucas C., Bayley R., Valli C.* (2001), Sociolinguistic Variation in American Sign Language. Washington, DC: Gallaudet University Press.
8. *Mathis T., Yule G.* (1994), Zero quotatives, Discourse Processes, Vol. 18, № 1, pp. 63–76.
9. *Pfau R., Steinbach M., Herrmann A.* (2016), A matter of complexity: subordination in sign languages, Boston: De Gruyter Mouton.
10. *Safar E., Glauert J.* (2012), Computer modelling, Sign Language: An International Handbook, pp. 1075–1101.
11. *Schlenker P.* (2017) Super monsters I: Attitude and Action Role Shift in sign language, Semantics and Pragmatics, Vol. 10, № 9.
12. *Schembri A., Johnson T.* (2012) Sociolinguistic aspects of variation and change, Sign Language: An International Handbook, pp. 788–816.
13. *Quer J.* (2011), Reporting and quoting in signed discourse, Quotation, Berlin: De Gruyter Mouton, pp. 277–302.

LANGUAGE PRODUCTION AND COMPREHENSION IN FACE-TO-FACE MULTICHANNEL COMMUNICATION[1]

Kibrik A. A. (aakibrik@gmail.com),
Fedorova O. V. (olga.fedorova@msu.ru)
Institute of Linguistics RAS and
Lomonosov Moscow State University, Moscow, Russia

Although language production and comprehension are parts of one and the same linguistic capacity, they have been studied separately for a long time. A key issue in the present day research is how the two processes are related, and whether transitions from thought to language and vice versa are accomplished by a single or two separate systems. Important progress in this area has been achieved in the field of psycho- and neurolinguistics; a brief review is provided in **Section 1**. In this paper we explore the production—comprehension relationship on the basis of our multichannel resource "Russian Pear Chats and Stories". In **Section 2** we describe this resource, including the stimulus material, data collection setup, participants and corpus size, and technical aspects. **Section 3** lays out two main theoretical notions: a model of face-to-face multichannel communication and a scheme of the production-comprehension interweaving in each interlocutor. In subsequent sections we discuss three case studies of production—comprehension relationships: relative contributions of kinetic channels to discourse understanding (**Section 4**), turn-taking and eye gaze (**Section 5**), and multichannel continuity (**Section 6**). The evidence of the multichannel corpus suggests a cognitive architecture that integrates language production and comprehension.

1. Language production and comprehension in psycho- and neurolinguistic studies

A well-known integrated model of production and comprehension is proposed by [Pickering and Garrod 2013][2]. They critically represent the traditional view as a "cognitive sandwich" (a term from [Hurley 2008]), in which action (including production) and perception (including comprehension) are separate and distinct (see e.g. [Dell 1986] for language production and [MacDonald et al. 1994] for language comprehension). In contrast to this view, Pickering and Garrod argue that production and comprehension are interwoven, and that such interweaving is what enables

[1] This study is supported by Russian Science Foundation (grant #14-18-03819 "Language as is: Russian multimodal discourse").

[2] For other integrated approaches see the CAPPUCCINO model by [McCauley and Christiansen 2011] and the "P-chain" framework by [Dell and Chang 2013], unifying the processes of comprehension, production and acquisition.

people to make predictions regarding themselves and others. Pickering and Garrod use the notion of forward modeling in action[3], grounded in computational neuroscience [e.g. Wolpert et al. 2003]. According to this notion, speakers employ "forward production" in predicting their upcoming utterances. Listeners also use forward production models and covertly imitate speakers to predict their production. As Pickering and Garrod put it, "the account helps explain the rapidity of production and comprehension and the remarkable fluency of dialogue" [2013: 346].

The model introduced in Pickering and Garrod 2013 suggests that prediction plays a central role in language production and comprehension. The recent paper "Prediction is Production: The missing link between language production and comprehension" [Martin et al. 2018] just published in "Nature" is testing this hypothesis. According to the authors, except some indirect support showing that language production skills and prediction are related ([Hintz et al. 2016]; [Federmeier et al. 2010]; [Huettig 2015]), there has been no direct evidence so far that the production system is necessary for prediction during comprehension. To test this hypothesis, [Martin et al. 2018] explore whether availability of the production system is indeed necessary for prediction during sentence comprehension. They compared three groups of participants reading Spanish sentences containing an expected vs. an unexpected NP such as *El rey llevaba en la cabeza* ***una corona / un sombrero*** *antigua/antiguo* 'The king wore on his head an old **crown** [Fem] / **hat** [Masc]'. Lexical prediction can be measured through event-related potential responses derived from electrophysiological recording during sentence reading: the less predictable a word is, the more negative is the N400 component. Participants of the first group were assigned a secondary verbal task preventing them from using their inner speech, while participants from two other groups had other secondary tasks, presumably language-unrelated. The expectation effect was reduced in the first group compared to two others groups. This finding demonstrates that preventing subvocal rehearsal of the verbal input during sentence reading hinders prediction in sentence comprehension.

Neuroimaging studies of language have typically focused on either production or comprehension of speech material such as syllables, words, or sentences. A study reported in Silbert et al. 2014 challenges this common practice, as well as the traditional assumption that the linguistic processes are primarily lateralized in the left hemisphere (see e.g. [Indefrey and Levelt 2004]; [Indefrey 2011]; but cf. [Jung-Beeman 2005]; [Hickok and Poeppel 2007] on language comprehension). [Silbert et al. 2014], looking at the production and comprehension of spontaneous narratives, identified all areas that are reliably activated in the brains of speakers telling a 15 minutes long narrative. Next, they identified areas that are reliably activated in the brains of listeners as they comprehended the same narrative. The results indicate that narrative production is not localized in the left hemisphere but recruits an extensive bilateral network, which overlaps extensively with the comprehension system. This study provides strong evidence for a close link between production and comprehension processes. Silbert et al. argue that "a shared neural mechanism supporting both production and

[3] Cf. a similar set of ideas in the early theory of motor control proposed by [Nikolai A. Bernstein 1967].

comprehension facilitates communication and underline the importance of studying comprehension and production within unified frameworks" [2014: E4687].

To sum up, a number of current experimentally oriented students of communication assume that production and comprehension are interwoven, that prediction plays a central role in language production and comprehension, and that brain networks involved in language production and comprehension strongly overlap. The reviewed studies are limited to the unimodal perspective, according to which language is a purely vocal phenomenon. In this study we put the discussion of the production—comprehension relationship in a broader context of multimodal/multichannel communication.

2. Russian Pear Chats and Stories

We explore the production—comprehension relationships on the basis of our new resource "Russian Pear Chats and Stories" [Kibrik 2018b; Kibrik and Fedorova 2018]. We have used the well known Pear Film [Chafe ed. 1980] as the stimulus material for collecting recordings and the state of the art equipment including three individual industrial cameras (100 fps) and two eyetrackers Tobii Glasses II (50 Hz).

We have developed a new procedure of data collection. Each session lasted for about one hour and involved four participants with fixed roles—the Narrator (N), the Commentator (C), the Reteller (R), and the Listener (L). Before recording began, the Narrator and the Commentator each watched the film on a personal computer, trying to memorize the plot as precisely as possible. Then the Narrator told the Reteller about the plot of the film; this is a monologic stage—*first telling*. During the subsequent, interactive, stage—*conversation*—the Commentator added details and corrected the Narrator's story where necessary, and the Reteller checked his/her understanding of the plot, asking questions to both interlocutors. Then the Listener joined the group and another monologic stage—*retelling*—followed, during which the Reteller was retelling the film to the Listener. Finally, the Listener wrote down the content of the film.

The resource includes 40 sessions, with 160 Russian native speakers aged 18–36, including 60 men and 100 women. The overall volume of the resource is 15 hours of recording and about 170 K words. Vocal (auditory) data are annotated using the Praat program (fon.hum.uva.nl/praat), in accordance with a scheme including temporal dynamics, segmentation into elementary discourse units (EDUs), absolute and filled pauses, accents, accelerated tempo, reduced pronunciation, lowered tonal register, etc. [Kibrik and Podlesskaya eds. 2009]. For the transcription of the kinetic (visual) data we used the annotation software ELAN (lat-mpi.eu/tools/elan). We annotated the following layers for facial/cephalic/manual/torso gestures: movements, movement chains, gestures, gesture chains, gesture phases, self-adaptors, postures, and posture changes. (See Litvinenko et al. 2017 for a more detailed description.) Gaze targets are coded as "surroundings" or "interlocutor", the latter further subdivided into "face", "hands", "torso", and "other". The minimal fixation duration is 100 ms, i.e. a participant' fixation on a target must last for at least 100 ms to be recognized as a gaze event [Fedorova 2017].

3. Face-to-face multichannel communication: Theoretical schemes

In face-to-face communication, interlocutors combine verbal structure, prosody, eye gaze, as well as facial, head, hand and torso gestures to produce integrated multichannel discourse. All of these communication channels are employed simultaneously and in conjunction with each other [Kress 2002; Kibrik 2010, 2018a,b; Müller et al. eds. 2014], see **Fig. 1**.

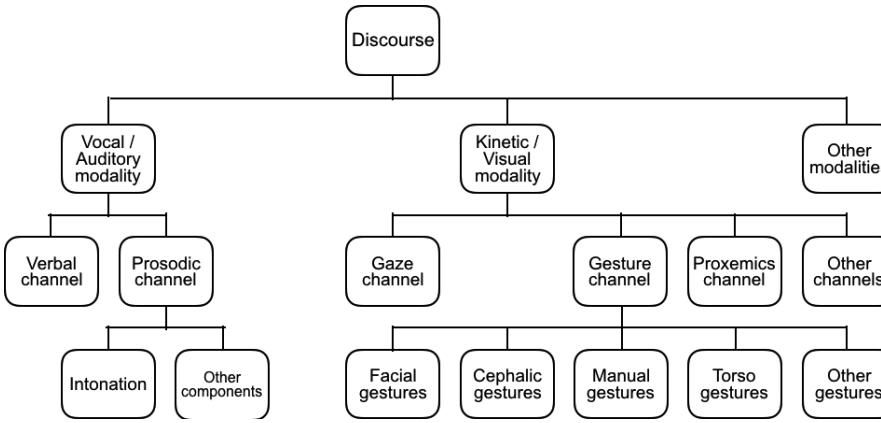


Fig. 1. Model of multichannel discourse

From the perspective of comprehension, one distinguishes the auditory and visual modalities; in terms of production, the same modalities can be dubbed vocal and kinetic.

The major change to the traditional notion of unimodal linguistic communication, necessitated by the bimodal (and multichannel) approach such as shown in **Fig. 1**, is the following. During the process of face-to-face communication each interlocutor performs the roles of addresser and addressee simultaneously. For example, a speaker, while producing speech at a given moment, simultaneously monitors the listener's kinetic behavior (nodding, gaze, and manual gesticulation). **Figure 2** depicts the production-comprehension ensemble.

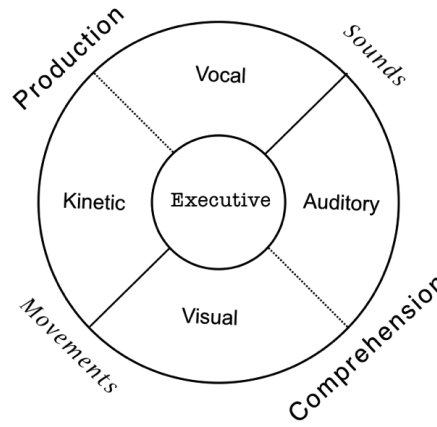


Fig. 2. Scheme of the production-comprehension interweaving in each interlocutor, taking part in face-to-face multichannel communication

The Executive is the central controlling component of the system (cf. similar executive components in theoretical models such as in [Baddeley 2007]; [Levelt 1989]; [McNeill 1992]). As in other models of cognitive processing, the Executive controls attentional processes and enables the system to selectively attend to some stimuli and ignore others.

Relying on the data of our resource, we discuss below three case studies of how processes of production and comprehension are interwoven in natural communication.

4. Case study 1. Kinetic channels: Relative contributions to discourse comprehension

In previous work we addressed the question of the relative contribution of various communication channels to the overall comprehension of spoken discourse. [Kibrik and Molchanova 2013] considered three communication channels employed in multichannel discourse in isolation: the verbal component, prosody, and kinetic-visual behavior. They found that all three channels play an important role in the overall process of conveying a message from a speaker to an addressee. They also found that participants had difficulties integrating the information from the visual and the prosodic channels, in the absence of the verbal channel. This suggests that in normal communication the verbal channel plays the role of an anchor to which the information from other channels is attached.

In [Kibrik and Molchanova 2013], the choice of a certain isolated communication channel was imposed upon experiment participants. In this study we observe how participants dynamically choose themselves which channel is most relevant at the given time from the point of view of their current communicative goals. Such choice is registered with the help of eyetrackers identifying attention allocation.

In Section 2 we mentioned three stages of communication events, analysed in the "Russian Pear Chats and Stories" corpus. The second of these stages, *conversation*, is interactive: all the interlocutors actively contribute vocal and kinetic material. The first and the third stages (*first telling* and *retelling*) are monologic: in each of them only one participant is actively talking, that is producing the vocal signal. But the roles are distributed differently: in the *first telling*, the main speaker is the Narrator, while in *retelling* the Reteller. This variation across the three stages allows us to compare the participants' oculomotor behavior. Specifically, we compare the visual attention[4] distribution in one participant: the Narrator. The analysis is based on three recordings: 04, 06 and 23.

Table 1. The distribution of Narrator's visual attention, *first telling* (summary duration, s and in %)

Recording #	Gaze target		Total
	face R	surroundings	
04	143.767 (61.2%)	91.205 (38.8%)	234.972 (100%)
06	47.908 (46.0%)	56.646 (54.0%)	104.554 (100%)
23	63.849 (38.3%)	103.032 (61.7%)	166.881 (100%)

As the data in Table 1 suggest, a primary speaker divides his/her visual attention exclusively between the primary listener's face (in this case, the Reteller) and the surroundings.

Table 2. The distribution of Narrator's visual attention, *conversation* (summary duration, s and in %)

Recording #	Gaze target					Total
	face R	face C	hands R	hands C	surroundings	
04	386.078 (65.5%)	40.077 (6.8%)	4.840 (0.8%)	0.320 (0.1%)	158.390 (26.8%)	589.705 (100%)
06	236.841 (46.7%)	95.216 (18.8%)	29.114 (5.7%)	29.074 (5.7%)	116.844 (23.1%)	507.089 (100%)
23	65.819 (20.9%)	178.228 (56.5%)	0.000 (0%)	1.380 (0.4%)	69.983 (22.2%)	315.41 (100%)

[4] It is generally recognized that attention and eye movements are closely related, even though the nature of this relationship is not yet fully understood; see e.g. [Smith and Schenk 2012].

As is clear from Table 2, when the Narrator is involved in multi-party discourse, watching the surroundings takes significantly[5] less time; the results for the Narrators in all three recordings are statistically indistinguishable. Apart from that, his/her attention is distributed between the interlocutors' faces and, to some extent, their hands. This apparently depends on the level of their activity in conversation.

Table 3. The distribution of Narrator's visual attention, *retelling* (summary duration, s and in %)

	Gaze target			
Recording #	face R	hands R	surroundings	Total
04	246.541 (74.2%)	54.826 (16.5%)	31.123 (9.3%)	332.490 (100%)
06	324.598 (75.5%)	57.914 (13.5%)	47.626 (11.0%)	430.138 (100%)
23	184.728 (77.2%)	12.697 (5.3%)	41.945 (17.5%)	239.370 (100%)

The evidence in Table 3 demonstrates that, while listening to the Reteller, the Narrator directs his/her gaze at the surroundings to a still lesser extent (the difference is significant for recordings 04 and 06). The vast majority of fixations are on one interlocutor: the Reteller, who is the primary speaker; the results in all three recordings are statistically indistinguishable. The Narrator's attention is distributed between the Reteller's face and, to a lesser extent, his/her hands.

We thus can conclude that a participants's visual attention is distributed in a systematic way, depending on his/her role as the primary speaker vs. an equal interlocutor vs. a listener. This distribution varies from one discourse stage to another and is sensitive to the participants' communicative activity.

5. Case study 2. Turn-taking and eye gaze

One of the chapters of the monograph [Grishina 2017] is titled "A grammar of gaze". In that chapter Elena A. Grishina explored the gaze direction of interlocutors at turn boundaries. Her analysis is based on the Russian movies ["Brilliantovaja ruka" The Diamond Arm (1968)] and ["Den' vyborov" Election Day (2007)]. Grishina found substantial differences in the speaker's gaze distribution at turn boundaries (her sample includes 527 instances). In particular, when a speaker directly addresses the listener (a "provoking" speech act), at the end of his/her turn s/he watches the listener, thus controlling the process of turn handover. If a speaker performs a neutral speech act, not demanding an immediate response from the listener, his/her gaze is usually directed at the surroundings.

[5] Here and below the statistical significance was analysed with the help of the chi-square test (Holm correction for multiple comparisons, $p < 0.01$).

We have tested this generalization against our data. In our pilot study, we looked at the conversation stage (8.5 min.) of recording 22 of the "Russian Pear Chats and Stories" corpus. Our material suggests two important differences from the Grishina's material. First, we analyse the talk and gaze of three interlocutors. Second, the behavior of two interlocutors was registered with the help of eyetrackers, which ensures a high precision of annotation. In the analysis we use the scores vocal transcript, prepared by Nikolay A. Korotaev. In the analysed conversation we identified 107 instances of turn boundaries, among these 32 were preceded by provoking turns and 75 by neutral turns. Table 4 illustrates an excerpt from the conversation's vocal transcript.

Table 4. An example of multichannel turn-taking

TimeS	TimeE	Narrator		Commentator		Reteller	
701,73						R-v076	Но не \очень симпатичная девочка¿
702,89	703,08	N-v329	\Нет,				
703,08		N-v330	она не \очень /симпатичная, ≈≈				
703,24				C-v261	(' 0.61) Не \знаю.		
	703,30						
703,30						R-v077	{laugh 8.32}
	704,53						
	704,59						
704,59	706,15			C-v262	/Вполне себе симпатичный		
706,48	707,15			C-v263	\Круглощёкий.		

In the excerpt shown in Table 4 the Narrator first watches the Reteller. At the beginning of her turn (N-v329) she moves her gaze to the surroundings and then, cued by the Commentator's voice, shifts her gaze at him. As for the Commentator, at the beginning of his turn (C-v261) he watches the Reteller, and then shifts his provoking gaze at the Reteller, waiting for her response.

The obtained results generally accord with Grishina's conclusions, but some important differences have been noted. First, the interlocutors' gaze is rarely directed at the surroundings; 75% of the time it is distributed between two partners in communication. Still, during those intervals when interlocutors' gaze is actually directed at the surroundings, 95% of the time this is in line with the Grishina's generalization stating that this happens in neutral speech acts. Second, the basic principle "Watch the interlocutor who has just started talking" is violated in certain instances (accounting for 12% of all instances); according to our current interpretation, this happens when a speaker assesses his/her own vocal contribution as being of low significance.

6. Case study 3. Multichannel continuity

The division of communication into three stages, including the *first telling*, the *conversation*, and the *retelling*, was originally seen as a technical procedure but later developed into a research issue in its own right. Multichannel communication is so organized that identifying boundaries between stages is rarely an easy task. Various channels suggest their own boundaries that do not have to coincide. Whereas the verbal, the prosodic, and the manual-gestural components are relative well coordinated (cf. Fedorova et al. 2016 on the degree of such coordination), the cephalic gestures, the facial gestures and eye movements regularly disturb coordination. In particular,

in the course of a vocal-manual pause, typical of a stage boundary, interlocutors usually produce signals that convey "turn handover". Such a signal is often a head turn, accompanied by particular facial movements, especially smile. Furthermore, the interlocutor's gaze frequently lags behind, remaining fixed on the other participant.

Consider an example from recording 35, specifically the boundary between the *first telling* and the *conversation*. Table 5 illustrates a five seconds excerpt that embraces a whole gamut of various vocal and kinetic actions performed by the interlocutors. A vocal expression is shown in line 1 of the Table. The behaviors listed in line 9 and further are clearly separable from that vocal expression, and we posit a boundary at the end of line 1, that is at 0.8 s from the beginning. In contrast, the behaviors listed in lines 2 to 8 all intersect that boundary.

Table 5. Recording #35, around the boundary between the *first telling* and the *conversation*

1.	N: vocal	Я закончила.	00.000–00.800
2.	R: manual	adaptor II[6]	00.000–05.100
3.	R: gaze	fixation	00.000–01.010
4.	N: gaze	fixation	00.000–01.620
5.	C: gaze	fixation	00.000–02.240
6.	N: cephalic	nod	00.150–01.120
7.	N: manual	adaptor II	00.250–05.100
8.	R: facial	smile	00.550–05.100
9.	R: cephalic	turn	00.960–01.800
10.	N: cephalic	turn	01.580–02.360
11.	C: cephalic	turn	02.280–02.570
12.	R: cephalic	nod	02.380–02.830
13.	C: cephalic	turn	03.380–04.010
14.	N: vocal	Ну?..	03.600–03.800
15.	C: vocal	Я не помню…	04.160–04.980
16.	C: manual	adaptor II	04.240–05.100

We can thus generalize that communication between interlocutors is not interrupted even for a fraction of a second. It is being supported by a network of channels. Work load is being swiftly and dynamically carried over from one channel to another. An interlocutor simultaneously functions as an addresser and an addressee.

7. Conclusion

Traditionally, production and comprehension are regarded as distinct processes. Some modern approaches, however, amend this dichotomy, proposing that production and comprehension are interwoven, and such interweaving is possible on the

[6] So-called "adaptors II" are minor movements without a clearly identifiable communicative function.

basis of prediction. Some recent studies on the functional neuroanatomy of language suggest that the brain networks involved in speaking and listening strongly overlap.

The evidence of multichannel communication also suggests a cognitive architecture that integrates language production and comprehension. As Pickering and Garrod said, "interlocutors must simultaneously produce their own contributions and comprehend the other's contribution. Clearly, an approach to language processing that assumes a temporal separation between production and comprehension cannot explain such behavior" [2013: 330]. Communicative actions of the interlocutors thus form a complex and heterogeneous network that must be credited a capability to involve simultaneous and multidirectional thought exchange.

References

1. *Baddeley A. D.* (2007), Working memory, thought, and action, Oxford: Oxford University Press.
2. *Bernstein N. A.* (1967), The co-ordination and regulation of movements, Oxford: Pergamon Press.
3. *Chafe W.* (ed.) (1980), The pear stories: Cognitive, cultural, and linguistic aspects of narrative production, Norwood: Ablex.
4. *Dell G. S.* (1986), A spreading-activation theory of retrieval in sentence production, Psychological Review, Vol. 93, pp. 283–321.
5. *Dell G. S., Chang F.* (2013), The P-chain: relating sentence production and its disorders to comprehension and acquisition, Philosophical Transactions of the Royal Society B: Biological Sciences, Vol. 369 (1634).
6. *Federmeier K. D., Kutas M., Schul R.* (2010), Age-related and individual differences in the use of prediction during language comprehension, Brain Language, Vol. 115 (3), pp. 149–161.
7. *Fedorova O. V.* (2017), Distribution of the interlocutors' visual attention in natural communication: 50 years later [Raspredeleniye zritel'nogo vnimaniya sobesednikov v estestvennoy kommunikatsii: 50 let spustya], E. V. Pechenkova, M. V. Falikman (eds.) Cognitive science in Moscow: new research. Proceedings of the conference [Kognitivnaya nauka v Moskve: novye issledovaniya. Materialy konferentsii]. Moscow: BukiVedi, IPPiP, pp. 370–375.
8. *Fedorova O. V., Kibrik A. A., Korotaev N. A., Litvinenko A. O., Nikolaeva Ju. V.* (2016), Temporal coordination between gestural and speech units in multimodal communication [Vremennaya koordinatsiya mezhdu zhestovymi i rechevymi edinitsami v mul'timodal'noy kommunikatsii], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii], RGGU, Moscow, pp. 159–170.
9. *Grishina E. A.* (2017), Russian gestures from a linguistic perspective [Russkaya zhestikulyatsiya s lingvisticheskoy tochki zreniya], Moscow: Jazyki slavyanskoy kul'tury.
10. *Hickok G., Poeppel D.* (2007), The cortical organization of speech processing, Nature Reviews Neurosciences, Vol. 8 (5), pp. 393–402.

11. *Hintz F., Meyer A. S., Huettig F.* (2016), Encouraging prediction during production facilitates subsequent comprehension: Evidence from interleaved object naming in sentence context and sentence reading, The Quarterly Journal of Experimental Psychology, Vol. 69 (6), pp. 1056–1063.
12. *Huettig F.* (2015), Four central questions about prediction in language processing, Brain Researcg, Vol. 1626, pp. 118–135.
13. *Hurley S.* (2008), The shared circuits model (SCM): How control, mirroring, and simulation can enable imitation, deliberation, and mindreading, Behavioral and Brain Sciences, Vol. 31 (01), pp. 1–22.
14. *Indefrey P.* (2011), The spatial and temporal signatures of word production components: a critical update, Frontiers in Psychology, Vol. 2, p. 255.
15. *Indefrey P., Levelt W. J.* (2004), The spatial and temporal signatures of word production components, Cognition, Vol. 92 (1–2), pp. 101–144.
16. *Jung-Beeman M.* (2005), Bilateral brain processes for comprehending natural language, Trends in Cognitive Science, Vol. 9 (11), pp. 512–518.
17. *Kibrik A. A.* (2010), Multimodal linguistics [Mul'timodal'naya lingvistika], Yu. I. Aleksandrov, V. D. Solov'yev (eds.), Cognitive studies [Kognitivnyye issledovaniya], Vol. IV, Institute of psychology, Moscow, pp. 134–152.
18. *Kibrik A. A.* (2018a), Russian multichannel discourse. Part I. Setting up the problem [Russkiy mul'tikanal'nyy diskurs. Chast' I. Postanovka problemy], Psikhologicheskiy zhurnal, Vol. 39 (1), pp. 70–80.
19. *Kibrik A. A.* (2018b), Russian multichannel discourse. Part II. Corpus development and avenues of research [Russkiy mul'tikanal'nyy diskurs. Chast' II. Razrabotka korpusa i napravleniya issledovaniy], Psikhologicheskiy zhurnal, Vol. 39 (2), pp. 78–89.
20. *Kibrik A. A., Fedorova O. V.* (2018), A "portrait" approach to multichannel discourse, Eleventh International Conference on Language Resources and Evaluation (LREC), Japan, 5–12 May 2018.
21. *Kibrik A. A., Molchanova N. B.* (2013), Channels of multimodal communication: Relative contributions to discourse understanding, M. Knauff, M. Pauen, N. Sebanz, I. Wachsmuth (Eds.), Proceedings of the 35th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society, pp. 2704–2709.
22. *Kibrik A. A., Podlesskaja V. I.* (eds.), (2009), Night Dream Stories: A corpus study of spoken Russian discourse [Rasskazy o snovideniyakh: korpusnoye issledovaniye russkogo ustnogo diskursa]. Moscow: Jazyki slavyanskikh kul'tur.
23. *Kress G.* (2002), The multimodal landscape of communication, Medien Journal, Vol. 4, pp. 4–19.
24. *Levelt W. J. M.* (1989), Speaking: From intention to articulation, MIT Press.
25. *Litvinenko A. O., Nikolaeva Ju. V., Kibrik A. A.* (2017), Annotation of Russian manual gestures: Theoretical and practical issues [Annotirovaniye russkikh manual'nykh zhestov: teoreticheskiye i prakticheskiye voprosy], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2017"], Moscow: RGGU, pp. 255–268.

26. *MacDonald M.C., Pearlmutter N. J., Seidenberg M. S.* (1994), The lexical nature of syntactic ambiguity resolution, Psychological Review, Vol. 101, pp. 676–703.
27. *Martin C. D., Branzi F. M., Bar M.* (2018), Prediction is production: The missing link between language production and comprehension, Nature, Vol. 8 (1079).
28. *McCauley S., Christiansen M. H.* (2011), Learning simple statistics for language comprehension and production: The CAPPUCCINO model, L. Carlson, C. Hölscher, T. Shipley (eds.), Proceedings of the 33rd annual conference of the cognitive science society, Austin, TX: Cognitive Science Society, pp. 1619–1624.
29. *McNeill D.* (1992), Hand and mind: What gestures reveal about thought, Chicago: University of Chicago Press.
30. *Müller C., Fricke E., Cienki A., McNeill D.* (eds.) (2014), Body—Language—Communication: An international handbook on multimodality in human interaction, Berlin: Mouton de Gruyter.
31. *Pickering M. J., Garrod S.* (2013), An integrated theory of language production and comprehension, Behavioral and Brain Sciences, Vol. 36 (04), pp. 329–347.
32. *Silbert L. J., Honey C. J., Simony E., Poeppel D., Hasson U.* (2014), Coupled neural systems underlie the production and comprehension of naturalistic narrative speech, Proceedings of the National Academy of Sciences of the United States, Vol. 111, pp. E4687–E4696.
33. *Smith D. T., Schenk T.* (2012), The premotor theory of attention: time to move on?, Neuropsychologia, Vol. 50, pp. 1104–1114.

CREATING A CORPUS OF SYNTACTIC CO-OCCURRENCES FOR RUSSIAN[1]

Klyshinsky E. S. (klyshinsky@mail.ru)
Keldysh IAM RAS, Moscow, Russia

Lukashevich N. Y. (natalukashevich@mail.ru),
Kobozeva I. M. (kobozeva@list.ru)
Moscow State University, Moscow, Russia

In the paper we discuss methods used to create CoSyCo, a corpus of syntactic co-occurrences, which provides information on syntactically related words in Russian. We describe a list of shallow parsing templates, which were used to collect data for CoSyCo. The paper includes an overview of the corpora collected for CoSyCo creation and an outline of how the noun 'virus' is used in its subcorpora as an example of the information which can be obtained from this online resource.

Keywords: corpora creation, shallow parsing, grammatically ambiguous text, words combinations, the Russian language

ОПЫТ СОЗДАНИЯ КОРПУСА СИНТАКСИЧЕСКИХ КОМБИНАЦИЙ РУССКОГО ЯЗЫКА

Клышинский Э. С. (klyshinsky@mail.ru)
ИПМ им. М. В. Келдыша РАН, Москва, Россия

Лукашевич Н. Ю. (natalukashevich@mail.ru),
Кобозева И. М. (kobozeva@list.ru)
МГУ им. М. В. Ломоносова, Москва, Россия

В данной статье дается основная информация о составе нового открытого ресурса — КоСиКо, корпуса синтаксических комбинаций, содержащего синтаксически связанные группы слов русского языка. Описывается состав текстовых корпусов, использованных для формирования КоСиКо, дается информация о шаблонах анализа текстов, использованных для извлечения информации. На примере употребления слова «вирус» с прилагательными показано, какого рода информацию можно получить из корпуса.

Ключевые слова: текст без снятой омонимии, поверхностный синтаксический анализ, создание корпуса, лексическая сочетаемость, русский язык

[1] The project is partially funded by RFH grant 15-04-12019.

1. CoSyCo: a Corpus of Syntactic Co-occurrences

In this paper we continue a series of works introducing a Corpus of Syntactic CoOccurrences (CoSyCo)[2]—a new resource providing information on word combinations in Russian.

It allows to get lists of word combinations together with examples of sentences in which they are used in real texts in the Internet giving information on word's co-occurrences, on syntactic relations between words.

We have already briefly outlined in (Klyshinsky, Lukashevich, 2017) the current state of affairs with regard to online resources providing similar information for Russian: one cannot say that there is a total lack of them. However, we believe that a freely accessible database which would be of a size big enough for various kinds of tasks, collected over huge untagged corpora with simple methods, offer a convenient interface and certain other important features is still to be designed [Klyshinsky, Lukashevich, 2017].

In this work we would like to focus on the structure and contents of CoSyCo database and to discuss methods used to create it.

2. CoSyCo structure

For this project we gathered data from open sources which we grouped in the following five subcorpora[3].

Table 1. CoSyCo subcorpora

	CoSyCo subcorpora	mln words	%
1.	News sites	1,400.9	8.07%
2.	IT news sites	142.4	0.82%
3.	Lib.rus.ec fiction collection	~15,000.0	86.38%
4.	Science sites	102.2	0.59%
5.	Wikipedia.ru texts (dump 01/05/2016)	~401.0	2.31%
6.	Russian Patents (http://www1.fips.ru/)	317.8	1.83%
	Total	~17,364.0	100.0%

[2] http://cosyco.ru/

[3] The fact that Librusec fiction collection by far outweighs all other subcorpora is to a great extent a result of technical issues (i.e. texts from which sites we managed to collect). Our intention was primarily to make a variety of text styles and genres available to a user. The importance of including texts which differ in style and genre into a corpus has been widely discussed in [Belikov et al, 2012], [Belikov et al, 2013], [Lukashevich et al, 2016]. Besides, which particular subcorpora size combination would make the corpus 'balanced' is not a trivial issue either. We definitely plan to increase the size of smaller subcorpora, but we believe that at the moment they can still be of help in a research as they are.

News sites included the following sources:

Table 2. News subcorpus in CoSyCo

News sites:	1,400.9 mln words	100%
lenta.ru	89.0	6.35%
RBK	66.0	4.71%
RIA Novosti	473.0	33.76%
Nezavisimaya gazeta	56.3	4.02%
Vzglyad	72.0	5.14%
Rossiyskaya gazeta	88.5	6.32%
Commersant	158.0	11.28%
Polit.ru	81.6	5.82%
Utro.ru	47.5	3.39%
Ibusiness	10.5	0.75%
Championat.com	1.2	0.09%
Moskovsky Komsomolets	72.1	5.15%
Gazeta.ru	78.4	5.60%
Komsomol'skaya Pravda	106.8	7.62%

IT news were taken from the following sources:

Table 3. IT news subcorpus in CoSyCo

IT news	142.4 mln words	100%
Membrana.ru	7.7	5.41%
CNews	43.6	30.62%
Computerra.ru	28.0	19.66%
Compulenta.ru	16.0	11.24%
PCWeek	23.3	16.36%
OSPNews	9.0	6.32%
Popular Mechanics	3.7	2.60%
NPlus1.ru	11.1	7.80%

Science sites covered a wide range of spheres and topics:

Table 4. Science subcorpus in CoSyCo

Science sites	102.2 mln words	100%
Childpsy.ru (dissertations)	12.6	12.33%
Civil Service Journal	1.8	1.76%
Delist.ru	8.9	8.71%
Dialogue conference	2.0	1.96%

Science sites	102.2 mln words	100%
Discollection.ru	11.3	11.06%
disser.aspirantura.spb.ru	1.6	1.57%
Geographic journals and books collection	4.3	4.21%
Musical journals	0.9	0.88%
Programming books collection	19.8	19.37%
Pu7.ru	18.5	18.10%
CAD and Graphics Journal	5.4	5.28%
Scientific Visualization Journal	0.2	0.20%
Information Security Journal	1.0	0.98%
Software Systems Journal	2.8	2.74%
Tomsk State University (TSU) Journal of Biology	1.0	0.98%
TSU Journal. Control, Computers and Informatics	0.6	0.59%
TSU Journal. Applied Discrete Mathematics	0.3	0.29%
TSU Journal	9.2	9.00%

3. Syntactic patterns used for data extraction

We have already partly described what methods and software were used to create CoSyCo database in [Klyshinsky et al., 2011], [Klyshinsky et al., 2016] and [Klyshinsky, Lukashevich, 2017]. One of the essential parts of CoSyCo project is a software tool for the extraction of syntactically connected words. This tool is written as a data-driven system that takes as input a template and a corpus and extracts combinations of a given format. In this part, we will discuss these templates in more detail.

It is known that in Russian some sequences of PoS-unambiguous words[4] can be considered as syntactically unambiguous without grammatical disambiguation. The structure of such sequences can be represented in the form of templates, which will help to identify whether the words in a phrase with the given structure are syntactically connected or not.

Experiments previously conducted on news corpora for seven European languages as described in [Klyshinsky et al, 2015] demonstrate that there are significant differences in the structure of homonymy/ambiguity: in Russian up to 50 % of words are unambiguous, and almost 80% of words are PoS-unambiguous (as compared to about 40% in English).

Another research in [Klyshinsky, 2017] focused on syntactic inversion in 33 languages. The study compared the number of left- and right-branching sentences for different types of syntactic links. The resulting figures demonstrated that the syntax of the Russian language is not so free as it may seem: Russian did not make it to the top ten languages with the free word order, coming in the middle of the rating.

[4] We understand a PoS-unambiguous word as a word with the same part of speech identified for every possible grammatical analysis.

These two points brought us to the general idea of the current research: if we take into account only very simple cases where it is easy to identify a syntactic relation between words (with no mistakes or with a negligible amount of them), and apply the corresponding templates to a very large corpus, with a comparatively high rate of PoS-unambiguous words and a relatively strict word order it should be possible to find most of possible combinations for a representative amount of words.

Certain points should be explained here. We did not plan to use various available tools for homonymy disambiguation in our work, because we wanted to avoid mistakes which they inevitably add. As for the existing search tools in tagged corpora (like Sketch Engine), they do not allow to work with non-disambiguated texts. Since we were particularly interested in this task, we needed to develop our own tool for it.

The next step was to formulate the templates for extracting syntactically connected words and check their work manually (for details see [Klyshinsky, Lukashevich 2017]).

Below, we will describe several templates of this kind which were used on the initial stages of the project.

I. Under a noun phrase (NP) in our work we understand a sequence of adjectives (possibly combined with optional adverbs) and a noun[5] which agree in gender, number and case. Obviously a noun phrase in Russian may contain various other elements, but we take into account only those which have such clear structure and, moreover, which contain only PoS-unambiguous words. A prepositional phrase (PP) is a noun phrase with a preceding preposition (as in (1)).

(1) PP = Prep + NP = Prep + (Adv) + (Adj^{+}) + N

Вероника повернулась, чтобы встретиться
'Veronika turned to look

с	***мягкими***	***зелеными***	***глазами.***
into	***soft***	***green***	***eyes.'***

The group of templates below (II–VII) helps to establish whether there is a syntactic connection between words in certain positions in a sentence. For this group of templates it is important that all words should be PoS-unambiguous, and that the NPs and PPs mentioned in the templates should be clearly separable from the context before and after them (e.g., by the beginning or the end of the sentence, the use of a PoS-unambiguous verb, a preposition, etc). The second condition is true not only for templates II–VII, but for VIII–X as well.

II. If a sentence starts with a single NP (as in (2a)) or a PP (as in (2b)) and such a phrase is followed by a single verb, then the noun in such a phrase and the verb are syntactically connected[6].

(2) a.

Российские	***аналитики***	***соглашаются***	*с тем, что …*
Russian	***analysts***	***agree***	*that…*

[5] Adjectives here should not be in a short or comparative form.

[6] It is important to note that we are not concerned about the direction of the connection here.

b. ***На*** ***<u>севере</u>*** ***<u>граничит</u>*** *с Латвией.*
In the ***<u>North</u>*** *(it)* ***<u>borders</u>*** *Latvia.*

III. The noun in the first NP or PP[7] which is used after a single verb is syntactically linked with this verb.

(3) *Новая технология* ***<u>предоставляет</u>*** ***опытным*** ***<u>пользователям</u>*** *расширенный набор возможностей печати.*
The new technology ***<u>offers</u>*** ***experienced*** ***<u>users</u>*** *a broader range of printing options.*

IV. The same conclusion as in II that a noun and a verb are syntactically linked can be made if the NP or PP is placed at the beginning of a subordinate clause which starts with a connector after a comma and if this NP or PP is followed by a single verb.

(4) *Блатт хотел, чтобы* ***<u>сезон</u>*** ***<u>завершился</u>*** *в начале мая.*
Blatt wanted that the ***<u>season</u>*** ***<u>be over</u>*** *in the beginning of May.*

V. An adverb placed between a preposition, noun, conjunction, or personal pronoun and an adjective is syntactically connected to this adjective.

(5) *Знаменитые эльфийские лучники*
<u>практически</u> ***<u>беспомощны</u>*** *при такой погоде.*
The famous elven archers are
<u>virtually</u> ***<u>helpless</u>*** *in this kind of weather.*

VI. If a participle is used before a noun in NP or PP (i.e. the position of the participle is typical for an adjective), then it is syntactically connected to the noun.

(6) ***<u>Рассматриваемая</u>*** ***<u>проблема</u>*** *находится на стыке дисциплин.*
The <u>investigated</u> ***<u>problem</u>*** *is at the intersection of several domains.*

VII. If a participle is used after NP or PP, is separated from it by a comma and agrees with the noun in this preceding NP or PP in gender, number and case, then the participle is syntactically connected with the preceding noun.

(7) *Системная* ***<u>интеграция,</u>*** ***<u>проводимая</u>*** *на заводе компании, …*
The system ***<u>integration</u>*** ***<u>performed</u>*** *at the plant of the company…*

In all templates above it was necessary that all words should be PoS-unambiguous. However, certain cases of PoS-ambiguity can be successfully resolved during analysis[8]. Templates VIII, IX and X show examples of this.

---

[7] There may be several noun or prepositional phrases after a verb, we are talking about the first of them.

[8] This is especially important for Russian, where a lot of nouns are derived from adjectives, so that they are ambiguous in every form (e.g. больной ‘ill / an ill person’).

VIII. If NP or PP includes a word, which is ambiguous between an adjective and a participle (as in (8a) or it is ambiguous between an adjective and a noun (as in (8b))[9], and if there is a PoS-unambiguous adjective in the same phrase then the ambiguous word should be considered an adjective.

(8) a. (Prep)NP = (Prep) + ?**Adj**/ Part + Adj + Noun
В Москве прошло вручение премии имени Елены Мухиной,
которой награждаются люди ***с*** ***ограниченными*** ***физическими*** ***способностями***.
limited -ADJ/PART
The ceremony of Elena Mukhina's award,
which is granted to people ***with*** ***limited*** ***physical*** ***abilities***, *took place in Moscow.*

b. (Prep)NP = (Prep) + ?**Adj**/ Noun + Adj + Noun
Прямая ***длинная*** ***линия*** *лезвия была скошена к концу.*
direct-ADJ/a line-NOUN
The ***direct*** ***long*** ***line*** *of the blade was slanted towards its end.*

IX. If NP or PP is at the end of the sentence and its last word is ambiguous between a noun and a verb (9a) or a noun and an adjective or participle (9b), then this last word in the phrase should be considered N. (The sequence should also meet the necessary criterion that in the resulting phrase the noun agrees with the preceding adjective(s) in its gender, number and case. The same applies to (9b).)

(9) IX a. …(Prep)NP = (Prep) + (Adj) + ?**Noun**/Verb.
a. *Он уставился* ***на*** ***лобовое*** ***стекло***.
glass-NOUN / flow down—PAST-SG-N
He stared at the ***front*** ***window***.
IX. b. …(Prep)NP = (Prep) + (Adj) + **?Noun**/Adj.
b. *Предстоит долгий путь* *до* ***финишной*** ***прямой***.
It is still a long way *to the* ***home*** ***straight***.
direct-ADJ/a line-NOUN

X. If NP or PP is followed by a(nother) PP and the last word of the first phrase is ambiguous between a noun and a verb (10a) or a noun and an adjective or participle (10b), this last word should be considered a noun. (Here the first NP or PP should be preceded by a verb, a punctuation mark, or the beginning of the sentence.)

(10) X a. [(Prep) +…+?Noun/V]PP/NP + PP/NP
a. *Он разбил* ***оконное*** ***стекло*** *в школьном коридоре.*
glass-NOUN / flow down -PAST -SG-N
He broke a ***window*** *pane* *in the school's passage way.*

[9] which are the most typical ambiguity cases for adjectives

X b. [(Prep) +…+?Noun/Adj/Part] PP/NP + PP/NP

b. ***Сводные*** ***<u>данные</u>*** *о значениях параметров …*
data-N / give-ADJ/Part-PL
The ***integrated*** ***<u>data</u>*** *on the parameters …*

4. Improving the results

In this section we will discuss the results obtained with the initial set of templates, and what steps had to be taken to improve them. (It was briefly mentioned in [Klyshinsky, Lukashevich, 2017]), here we will try to go into more detail.)

When we assessed how complete the database of combinations was, we found that a certain part of vocabulary was missing. While checking why this happened, we saw that at least one reason was that words which are grammatically ambiguous in all their forms in Russian (e.g. ученый is ambiguous between a noun ‹a scientist› and an adjective ‘learned, academic’ in every form) were disregarded during processing, They proved to be so frequent, that this dropped the amount of identified nouns and adjectives down.

To avoid this, we had to lift certain restrictions in several templates—we had to allow words ambiguous between a noun and an adjective in templates I and IX[10]. (The resulting templates are I*, IXa*, and IXb* respectively).

I* PP = Prep + NP = Prep + (Adv) + (Adj+) + **<u>?Adj/ Noun</u>**

IXa* …(Prep)NP = (Prep) + **<u>?Adj/ Noun</u>** + <u>?Noun</u>/Verb

IXb* …(Prep)NP = (Prep)+ **<u>?Adj/ Noun</u>** + <u>?Noun</u>/Adj.

We also added a new template which identified verbs from short forms of participles and established a link between such a verb and a noun in the noun phrase.

IX. If a participle in a short form is followed by NP or PP, then it is syntactically linked with the noun in NP or PP, and the same holds true for its producing verb.

Similarly, if NP or PP at the beginning of the sentence is followed by a participle in a short form, the same conclusions can be made.

(11) a. *Вырезки не были* ***<u>разложены</u>*** ***в хронологическом*** ***<u>порядке</u>***.
The cuttings were not ***<u>placed</u>*** ***in chronological*** ***<u>order</u>***.

b. ***Личный*** ***<u>состав</u>*** ***<u>размещен</u>*** *в закрытом городке.*
The military ***<u>personnel</u>*** ***<u>was placed</u>*** *in a restricted-access town.*

[10] This technically meant that we had to “soften” our initial position that only PoS-unambiguous words should be taken into account. The table below shows that these changes significantly improved the figures in CoSyCo database. This increase in figures also allows to indirectly assess the relative percentage of words ambiguous between a noun and an adjective. We believe all of this to be important, that is why we deliberately give a detailed account of the course of work instead of simply showing the current set of templates.

Another change was an additional set of conditions in several templates. Template II, for example, will also hold true if the requested group is used after a punctuation mark, as in (2c):

(2) c. *Необходимо тестирование 60% программ,*
считают ***эксперты*** *Ассоциации.*
It is necessary to test 60% of software,
believe *Association* ***experts****.*

The table below shows the effect such amendments had on the figures in the database.

Combination	Lemma combinations, mln		Token combinations, mln		Total occurences, mln	
	old	new	old	new	old	new
noun+adj	12.1	18.3	25.5	39.8	383	746
verb+prep+noun	29.2	33.4	53.5	60.3	349	412
participle+noun	3.1		5.1		28.1	
participle+prep+noun	1.2		1.8		4.3	

Table 5. CoSyCo database before and after amending the templates

Combination	nouns		adjective		verbs	
	old	new	old	new	old	new
noun+adj	67,000	71,000	41,000	42,000		
verb+prep+noun	73,000	73,000			28,000	28,000
participle+noun	52,000				20,000	
participle+prep+noun	40,000				15,000	

We compared the vocabulary extracted from CoSyCo database with the one obtained from SynTagRus (for details see [Klyshinky, Lukashevich, 2017]) and with the dictionary of I-RU bigrams from the database of Collocations Colligations Corpora [Kormacheva et al, 2016]. We found out that the vocabularies of the latter two resources differ significantly. For most frequent words (with frequencies over 1,000) the differences were between 1% and 4%. However, for words with lower occurrence figures (over 10) the differences were between 30% and 70%, with CoSyCo vocabulary being more complete. Comparison results are shown in **Table 6**.

Table 6. Comparison of I-RU and CoSyCo vocabularies

Part of Speech	Frequency	I-Ru		CoSyCo	
		Not found in CoSyCo	Total	Not found in I-Ru	Total
Noun	>1,000	226 (4,3%)	5,229	523 (3,1%)	16,881
	>500	534 (6,4%)	8,376	1,333 (6,0%)	22,209
	>100	3,887 (18,4%)	21,122	7,987 (21,7%)	36,866
	>10	30,298 (49,1%)	61,720	27,102 (46,3%)	58,524
Adjective	>1,000	10 (0,6%)	1,677	4,138 (30,3%)	13,635
	>500	22 (0,8%)	2,728	6,890 (41,2%)	16,705
	>100	405 (6,4%)	6,312	14,390 (58,8%)	24,481
	>10	4,014 (28,3%)	14,192	23,026 (69,4%)	33,194
Verb	>1,000	21 (0,9%)	2,291	197 (2,0%)	9,975
	>500	56 (1,6%)	3,601	725 (6,1%)	11,975
	>100	426 (5,2%)	8,153	4,073 (24,6%)	16,561
	>10	3,670 (22,5%)	16,291	10,598 (45,6%)	23,219

To assess the recall for identified word combinations we did the following. We applied a "weak" template according to which any two words which are Adj + Noun are syntactically linked. Such a template will definitely bring many false connections, but the most frequent ones should presumably be correct. We analyzed 10,000 most frequent combinations obtained with the help of this "weak" templates. We found only 159 nouns (about 1.5%) for which in CoSyCo database there were less than 75% adjectives identified with the "weak" pattern. The links were mostly missing when there was a mistake (e.g. one of the words did not belong to the requested part of speech in the context). We checked adjectives with frequencies higher than 10, and for about 2,400 nouns we did not manage to find only 1% of such adjectives, whereas for 23 nouns over 5% of such adjectives were missing. This usually meant that the omitted links were in the least frequent part of the list.

5. Word combination types on the site

At the moment a user of the site can find data on the following types of combinations (the relevant type can be selected from the left side menu of the screen):

- **verb**+preposition+noun (***арендовать*** *у компании 'rent from a company'*)[11],
- **noun**+adjective (*компьютерный* ***вирус*** *'computer virus'*),
- **noun**+participle (*созданный* ***имидж*** *‹created image›*),
- **participle**+preposition+noun (***арендованный*** *у компании 'rented from a company'*),

[11] In the title, the head constituent for the combination is highlighted in bold. The order of the elements in the title does not necessarily coincide with the typical word order in sentences with such word combinations. This was done on purpose to keep the logic so that it helps to find the relevant section for a combination regardless of the real word order in the sentence.

- **adjective**+adverb (*очень* ***амбициозный*** *'very ambitious'*).

Fig. 1 shows what the search page of the site looks like.

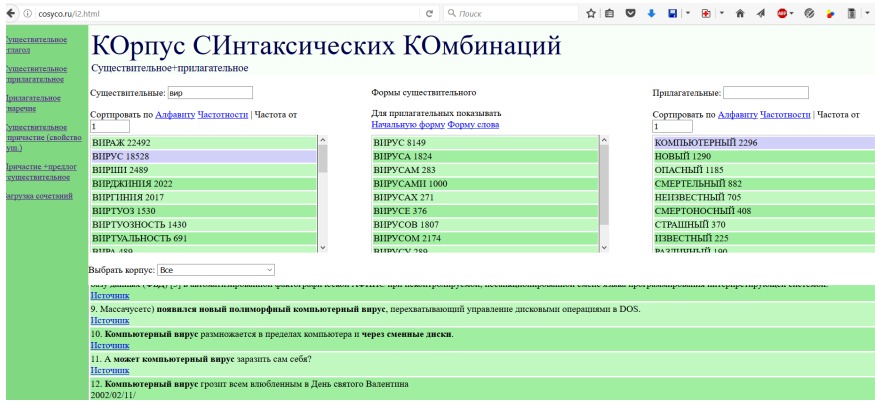


Fig. 1. CoSyCo search page showing examples for the word combination *компьютерный вирус* 'computer virus'

On the search page for the verb+preposition+noun and participle+preposition+noun sections, the left column lists verbs, the head constituent for the verb+preposition+noun combination and a derivational basis for the participle in participle+preposition+noun combination. In the middle column, a user can choose a preposition from the list of options that combine with the chosen verb, and in the right column s/he immediately gets a list of nouns which were found in the texts (=can be combined) with this verb+preposition and a list of sentences containing this expression in the lower part of the screen[12]. The resulting lists of words can be sorted by frequency or alphabetically.

In a similar way for the noun+adjective and noun+participle sections the left column gives a list of nouns. A user can choose a word form of the selected noun in the middle column and in the left column s/he sees a list of adjectives or participles (respectively) which were found in real texts as modifiers of the selected noun.

An important feature of CoSyCo is that it is possible to choose the source of examples with the expression in question to be shown on the screen. A user can leave the default "All" option on or can select one of the five subcorpora from a drop-down list; then the list of example sentences shown contains only those from the selected subcorpora. Although the example usually includes one sentence, it is possible to have a look at a broader context with the help of a link to the source text placed after the example.

[12] The figure after the verb shows how often the word is found in the whole corpus. The figures after words in the middle and right columns show absolute frequencies of respective word combinations in the whole corpus. Unfortunately, the problem of duplicates in the corpus is still being resolved, so the figures currently on the screen are not accurate.

6. Adjective + NOUN combinations across CoSyCo subcorpora and other resources

To try comparing the output of CoSyCo with that of existing resources of similar size we took the lists of most frequent adjectives used with the noun *вирус*'virus' in CoSyCo, RuTenTen and GICR.

The table below shows top ten most frequent adjectives used with this noun in the average CoSyCo collection with the figures for the same words in RuTenTen and GICR[13]. For each word "a" column contains absolute co-occurrence figures for the combination of this adjective with *вирус*'virus' in this corpus; "b" column shows how often this pair is found as compared to the total number of any adjective+ *вирус* 'virus' combinations in the corpus.

Table 7. Adjectives+ *вирус* 'virus' in CoSyCo, RuTenTen and GICR

	CoSyCo		RuTenTen		GICR	
	A	b	a	B	a	b
	42,321		95,040		25,267	
КОМПЬЮТЕРНЫЙ 'computer'	5,331	0.125965	12,257	0.128967	2,612	0.1033759
НОВЫЙ 'new'	3,030	0.071595	9,483	0.099779	2,346	0.0928483
ОПАСНЫЙ 'dangerous'	2,628	0.062096	5,237	0.055103	1,676	0.0663316
СМЕРТЕЛЬНЫЙ 'deadly'	1,899	0.044871	1,934	0.020349	823	0.0325721
НЕИЗВЕСТНЫЙ 'unknown'	1,195	0.028236	2,054	0.021612	442	0.0174931
СМЕРТОНОСНЫЙ 'lethal'	893	0.021100	718	0.007555	236	0.0093403
СТРАШНЫЙ 'dreadful'	589	0.013917	1,350	0.014205	709	0.0280603
ИЗВЕСТНЫЙ 'known'	396	0.009350	1,717	0.018066	143	0.0056596
ОБЫЧНЫЙ 'ordinary'	329	0.007773	576	0.006061	121	0.0047889

The figures show that in general the lists of adjectives are rather similar, and the variations in their frequencies may be explained by differences in the corpus structure, the style and genre differences of texts constituting them.

We also analyzed lists of top 100 most frequent adjectives used with this noun from the point of view of semantic classes which could be identified there. The tables

[13] For GICR we analyzed only data from three out of four available segments—news, Zhurnalny zal and LiveJournal. It was not clear beforehand what picture VKontakte texts would give, so we decided not to include them without prior research.

below show data for two groups which proved to be most frequent in virtually all segments in CoSyCo and GICR.

The first group can be identified as describing various features of a computer virus. Table 8 shows that adjectives from this group penetrate most of modern text styles and genres, with the adjective *komp'uterny*'computer' outweighing all other words in this group.

Table 8. 'COMPUTER' group of adjectives used with *вирус* 'virus'

	Komp'uterny 'computer'		*pochtovy* 'mail'		*mobilny* 'mobile'		total for group	
Source	a	b	a	b	a	b	a	b
CoSyCo news	1,121	13.18%	7	0.08%	115	1.35%	1,365	16.05%
CoSyCo compnews	990	24.93%	130	3.27%	98	2.47%	1,441	36.29%
CoSyCo Librusec	2,869	10.99%	(24)	0.09%	(29)	0.11%	3,538	20.05%
CoSyCo Wiki	276	25.72%	4	0.37%	7	0.65%	334	31.13%
CoSyCo science	50	11.99%	1	0.24%	5	1.20%	73	17.50%
GICR news	432	11.11%	24	0.62%	18	0.46%	517	13.30%
GICR zhurzal	45	11.57%	2	0.51%	—	—	54	13.88%
GICR Livejournal	2,135	10.31%	34	0.16%	78	0.37%	2,687	12.80%

The second group includes various adjectives united by the component 'know' in their lexical meaning. These words are also widely used to characterize a virus in every segment.

Table 9. 'KNOWN' group of adjectives used with *вирус* 'virus'

	novy 'new'		*neizvestny* 'unknown'		*izvestny* 'known'		total for group	
Source	a	b	a	b	a	b	a	b
CoSyCo news	949	11.16%	187	2.19%	57	0.67%	1,245	14.64%
CoSyCo compnews	621	15.64%	130	3.27%	117	2.95%	891	22.44%
CoSyCo Libr	1,417	5.43%	839	3.21%	183	0.70%	2,883	11.04%
CoSyCo Wiki	34	3.17%	37	3.45%	23	2.14%	107	9.97%
CoSyCo science	2	0.48%	1	0.24%	5	1.20%	9	2.16%
GICR news	692	17.80%	66	1.70%	32	0.82%	809	20.81%
GICR zhurzal	18	4.63%	14	3.60%	—	—	39	10.03%
GICR ljournal	1,636	7.79%	362	1.72%	111	0.53%	2,296	10.94%

Tables 8 and 9 show that results are quite comparable and in general both groups demonstrate similar tendencies in respective segments of different sources.

7. Conclusion

In the paper we have described the text corpora and methods used to create CoSyCo, a corpus of syntactic co-occurrences which provides information on syntactically related words in Russian. Currently there is a lot of room for improvement: there is a need to address deduplication issues, to increase the number of word combinations available on the site, as well as to improve its interface and the quality of output. We also understand the necessity to conduct a thorough comparison of the output of CoSyCo with that of the existing resources of similar size, and intend to do this in the nearest future.

References

1. *Belikov V., Selegey V., Sharoff S.* (2012), Preliminary considerations towards developing the General Internet Corpus of Russian [Prolegomeny k proektu General'nogo internet-korpusa russkogo yazyka], Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog" 2012" [Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Po materialam ezhegodnoi Mezhdunarodnoii Konferentsii "Dialog 2012"], Bekasovo, vol. 1, pp. 37–49.
2. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.* (2013), Corpus as language: from scalability to register variation [Korpus kak yazyk: ot masshtabiruemosti k differentsialnoi polnote] Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialog" (2013) [Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Po materialam ezhegodnoi Mezhdunarodnoii Konferentsii "Dialog" (2013)], Bekasovo, vol. 1, pp. 83–96.
3. *Klyshinsky E., Kochetkova N., Litvinov M., Maximov V.* (2011), Method of POS-disambiguation using information about words co-occurrence (for Russian), Proceedings of the annual meeting of the Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL), Hamburg, pp. 191–195.
4. *Klyshinsky E., Ermakov P., Lukashevich N., Karpik O.* (2016) The Corpus of Syntactic Co-occurences: the First Glance, in Proc. of the Fifth International Conference on Analysis of Images, Social Networks and Texts (AIST 2016), pp. 85–90.
5. *Klyshinsky E.*(2017) The Freedom of the Russian Syntax is Slightly Exaggerated, in Proc. of Workshop on New Information Technologies in Automated Systems, pp. 112–116.
6. *Klyshinsky E., Lukashevich N.* (2017) Corpus of Syntactic Co-Occurrences: A Delayed Promise, Artificial Intelligence and Natural Language: 6th Conference, AINL 2017, St. Petersburg, pp. 121–131.
7. *Kormacheva D., Pivovarova L., Kopotev M.* (2014), Automatic Collocation Extraction and Classification of Automatically Obtained Bigrams, in Proceedings of Workshop on Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations (CCLCC 2014), pp. 27–33 .
8. *Lukashevich N., Klyshinky E., Kobozeva I.* (2016), Lexical research in Russian: are modern corpora flexible enough?, Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference "Dialog" (2016) [Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Po materialam ezhegodnoi Mezhdunarodnoi Konferentsii "Dialog" (2016)], Moscow, pp. 385–397.

LEARNING WORD EMBEDDINGS FOR LOW RESOURCE LANGUAGES: THE CASE OF BURYAT

Konovalov V. P. (vaskonov@yahoo.com),
Tumunbayarova Z. B. (zhargal@zabgu.ru)
Transbaikal State University, Chita, Russia

Word-vector representations have been extensively studied for rich resource languages with large text datasets. However, only a few studies analyze semantic representations of low resource languages, when only small corpus is available. In this study we introduce a methodology and compare techniques to learn semantic representations of low resource languages. The proposed methodology consists of defining accurate preprocessing steps, applying language-independent stemmer and learning word-vector representations. In addition, we propose a simple word embeddings evaluation scheme that can be easily adapted to any language. By using this methodology we learn word-vector representations for Buryat language. In order to promote further research we make the source code and the resulting word embeddings corpus publicly available.

Keywords: word2vec, word-embeddings, SVD, PMI, GloVe

ВЕКТОРНОЕ ПРЕДСТАВЛЕНИЕ СЛОВ ДЛЯ МАЛОРЕСУРСНЫХ ЯЗЫКОВ: НА ПРИМЕРЕ БУРЯТСКОГО ЯЗЫКА

Коновалов В. П. (vaskonov@yahoo.com),
Тумунбаярова Ж. Б. (zhargal@zabgu.ru)
Забайкальский Государственный
Университет, Чита, Россия

1. Introduction

Using word embeddings is a standard practice in NLP systems, both in shallow and deep architectures [Goldberg, 2016]. Word embeddings exploits statistical techniques to embed words in a vector space. In this space, words with similar meanings tend to be located close to each other. These techniques are based on the Harris distributional hypothesis [Harris, 1954], which says that words in similar contexts have similar meanings. This statement provides a framework to use semantic relationship between words. Word embeddings has been used in a wide variety of applications such as query expansion [Chen & Chen, 2007], building bilingual comparable corpora [Zhu, Li, Chen, & Yang, 2013], clustering [Di Marco & Navigli, 2013].

Classical count-based methods such as PMI matrices and SVD factorization were very popular to represent words as vectors. However, recently word2vec approach has been proposed to represent words as dense vectors by applying neural embedding methods [Mikolov, Chen, Corrado, & Dean, 2013]. The neural embedding methods are particularly computationally-efficient predictive model for learning word embeddings. It comes in two types, the Continuous Bag-of-Words model (CBOW) and the Skip-Grams with Negative Sampling model (SGNS). In this work we answer the question of which model to choose for low resource languages when only small amounts of data are available.

Word normalization is an important data preprocessing step for learning word-vector representations. It improves the vectors quality by reducing language variability. In order to reduce words to a common base form, most stemmers use the extensive set of linguistic rules developed for specific language [Porter, 1980]. Usually low resource languages lack rule-based stemmers. In this work, we examine to what extend language independent stemmer can improve word embeddings quality when only small training data is available.

In addition, we suggest a new scheme for word vector evaluation. We aim to develop a method that can address the common shortcomings mentioned in [Hill, Reichart, & Korhonen, 2016], at the same time this methods can be easily reproducible for any language.

The remainder of the paper is organized as follows. Section 2 gives information on Buryat language. Then Section 3 provides an overview of the methods for building word-vector representations. Section 4 describes language independent stemming approach. Section 5 specifies the experimental setup and describes the evaluation methodology. Finally, Section 6 provides results and comparisons of various word embedding techniques.

2. Buryat language

Buryat language is one of the Mongolic languages. The majority of Buryat speakers live in Russia along the northern border of Mongolia where it is an official language in the Buryat Republic, Ust-Orda Buryatia and Aga Buryatia. According to the Russian census of 2002, there are 353,113 native speakers in Russia. In addition, there are at least 100,000 native speakers in Mongolia and the People's Republic of China as well. There are regularly published Buryat newspapers, journals, books, films, television and radio

programs, however, according to UNESCO report, Buryat is considered to be an endangered language and at risk of disappearing [Janhunen, 2006]. Implementation of NLP tools is crucial for language preservation and development. The first syntactic treebank for Buryat language based on the Universal Dependencies was developed in [Badmaeva & Francis, 2017]. Buryat language has seven cases and two numbers. Its alphabet is based on the general Cyrillic scripts with three additional letters.

3. Background

There are two major word-vector representation methods: the count-based methods (PMI matrix, SVD factorization) and the neural embeddings methods (SGNS, CBOW).

The previous results suggest that the new embedding methods consistently outperform the traditional methods by a non-trivial margin on many similarity-oriented tasks [Baroni, Bernardi, & Zamparelli, 2014]. However this result was reported only for rich resources languages [Altszyler, Sigman, Ribeiro, & Slezak, 2016]. For example, the model used in [Baroni, Bernardi, & Zamparelli, 2014] was trained on 2.8 billion tokens constructed by concatenating ukWaC[1], the English Wikipedia[2] and the British National Corpus [3]. Unfortunately such big resources are not available for many languages, including Buryat language.

We strictly follow the notation that was used in [Levy, Goldberg, & Dagan, 2015], where $w \in V_W$ is collection of words and $c \in V_C$ their contexts, and V_W and V_C are the word and context vocabularies. The collection of observed word-context pairs Is D. The number of times the pair $(w,)$ appears in D is denoted as $\#(w,c)$. Then, $\#(w) = \sum_{c' \in V_C} \#(w,c')$ and $\#(c) = \sum_{w' \in V_W} \#(w',c)$ are the number of times w and c occurred in D, respectively.

When words and contexts are embedded in a space of d dimensions, each word $w \in V_W$ is represented as a vector $\vec{w} \in \mathbb{R}^d$ and each context $c \in V_C$ is represented as a vector $\vec{c} \in \mathbb{R}^d$.

In this work we focused on fixed-window bag-of-words contexts, where D is obtained by using a corpus $w_1, w_2, \ldots, w_n$ and defining the contexts of word w_i as the words surrounding it in an L-sized window $w_{i-L}, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{i+L}$.

Section 3.1 describes the traditional pointwise mutual information method, which follows by singular values decomposition method in the Section 3.2. In addition, we examined word embeddings learned by GloVe, SGNS and CBOW methods. Due to space limitations we omitted their description here. The GloVe (Global vectors for word representation) method was described in [Pennington, Socher, & Manning, 2014]. The neural based methods SGNS and CBOW were described in [Mikolov, Chen, Corrado, & Dean, 2013].

---

1 http://wacky.sslmit.unibo.it

2 http://en.wikipedia.org

3 http://www.natcorp.ox.ac.uk

3.1. Pointwise mutual information (PMI)

Traditionally, the word-vector representations can be achieved by constructing a high-dimensional sparse matrix M, where each row represents a word w in the vocabulary V_W and each column a context $\in V$. The cell value M_{ij} represents the association between the word w_i and the context c_j. Pointwise mutual information (PMI) is a traditional metric to measure this association [Church & Hanks, 1990].

$$PMI(w,c) = \log \frac{P(w,c)}{P(w)P(c)} \tag{1}$$

Good collocation pairs have high PMI because the probability of co-occurrence is only slightly lower than the probabilities of occurrence of each word. Conversely, a pair of words whose probabilities of occurrence are considerably higher than their probability of co-occurrence gets a small PMI score.

$PMI\ (w,c) = -\infty$ if the number of co-occurrence of w and c equals 0. In order to address this, positive PMI (PPMI) is used, in which all negative values are replaced by 0.

It is well-known that PMI (and PPMI) suffers from its bias towards infrequent events [Turney & Pantel, 2010]. A rare context c that co-occurred with a word w even once will often lead to relatively high PMI score because (c), which is in PMI's denominator, is very small. This causes a situation in which the most associated contexts with w are often very rare words.

This problem can be addressed by smoothing variation of PMI. Smoothing context distribution increases the probability of sampling a rare context ($P\ (c) > P\ (c)$ when c is rare), which reduces the PMI of (w,c) for any w co-occurring with the rare context c. The smoothed PMI is very effective and consistently improves performance across different tasks and methods [Levy, Goldberg, & Dagan, 2015].

$$PMI_\alpha(w,c) = \log \frac{P(w,c)}{P(w)P_\alpha(c)} \tag{2}$$

$$P_\alpha(c) = \frac{\#(c)^\alpha}{\sum_c \#(c)^\alpha} \tag{3}$$

3.2. Singular Value Decomposition (SVD)

The dense low-dimensional vectors can be obtained by applying truncated Singular Value Decomposition (SVD) [Eckart & Young, 1936]. Formally, SVD of matrix M is factorization of the form $U \cdot \Sigma \cdot V^T$, where U and V are orthonormal and Σ is a diagonal matrix of eigenvalues in decreasing order. We obtain $M_d = U_d \cdot \Sigma_d \cdot V_d^T$ by keeping only top d elements of Σ. The high dimensional spare word-vector representations of matrix M can be substituted by low dimensional dense vectors of W_{SVD} and C_{SVD} that represent words and contexts respectively.

$$W_{SVD} = U_d \cdot \Sigma_d,\ C_{SVD} = V_d \tag{4}$$

However, word-vector representations of W_{SVD} are not necessary the optimal for semantic tasks. It was shown that weighting the eigenvalues matrix Σ_d can have a significant effect on the performance [Levy, Goldberg, & Dagan, 2015].

$$W_{SVD}^p = U_d \cdot \Sigma_d^p \tag{5}$$

4. Word Normalization

Identifying the original forms of words is important for natural language processing applications. The goal of normalization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form, for instance

$$am, are, is \rightarrow be$$
$$car, cars, car's \rightarrow car$$

Normalization can involve either lemmatization or stemming.

Stemming usually refers to a crude heuristic process that chops off the ends of words, and removal of derivational affixes. As result of the process we get a stem that does not have to be a proper word. Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma. Therefore stemmers are much simpler, smaller and usually faster than lemmatizers, and for many applications their results are good enough.

Usually low resource languages lack NLP tools like stemmer/lemmatizer. There is no normalizer for Buryat language, however several techniques were proposed for Mongolian language [Fujii & Chimeddorj, 2012]. It is extremely important to develop normalizer for low resource language in order to alleviate language variability.

4.1. Yet Another Suffix Striper (YASS)

YASS is a statistical corpus-based stemmer that does not rely on linguistic expertise. It stems by clustering a lexicon without any linguistic input. Its performance is comparable to that obtained using standard rule-based stemmers such as Porter's. Information retrieval experiments done on English, French and Bengali datasets found YASS very effective [Majumder, et al., 2007].

The clusters are created using hierarchical approach and distance measures. Four distance functions D_1, D_2, D_3, D_4 were proposed. The main intuition behind defining these distances was to reward long matching prefixes, and to penalize an early mismatch. The D_3 distance function was found to be the most effective, so we focused on D_3 solely [Majumder, et al., 2007].

If the strings X and Y are of unequal lengths we pad the shorter string with null characters to make the strings lengths equal. The distance D_3 between two strings $X = x_0 x_1 \ldots x_n$ and $Y = y_0 y_1 \ldots y_n$ is as following

$$D_3(X,Y) = \begin{cases} \dfrac{n-m+1}{m} \displaystyle\sum_{i=m}^{n} \frac{1}{2^{i-m}} \; if \; m > 0 \\ \infty \; otherwise \end{cases} \tag{6}$$

where m denotes the position of the first mismatch between X and Y.

The distance function defined above is used to compose a distance matrix. Then the distance matrix is used to cluster words. Each cluster is expected to represent morphological variants of a single root word. The words within a cluster are stemmed to the "central" word in that cluster.

Three variants of hierarchical clustering algorithms were tested, namely, single-linkage, average-linkage and complete-linkage. In single-linkage clustering the similarity between two clusters is the maximal similarity between any two members of the groups. Complete-linkage clustering is similar to single-linkage, but instead of maximal similarity, it considers the minimal similarity between any two members as a clusters similarity. In average-linkage clustering the similarity between two clusters is the mean similarity between members of different clusters [Jain, Murty, & Flynn, 1999].

5. Experimental Setup

Section 5.1 describes the Buryat Wikipedia corpus. Section 5.2 specifies the methods that were used to learn the word-vector representations. Finally, the evaluation scheme is defined in Section 5.3.

5.1. Corpus

The models were trained on the Buryat Wikipedia, which consist of 1381 articles (each one is more than 50 words long). The articles were lower-cased and non-textual were removed. In addition, we excluded all words that contain non-Buryat characters. As result the corpus contains 406715 words (64403 unique words).

5.2. Training Embeddings

The models were derived using windows of 2, 5, 10 tokens to each side of the focus word. For every window size we calculated PMI[4] word representations and we learned a 50, 100, 500-dimensional representations with SVD, SGNS, CBOW and GloVe methods.

5.3. Evaluation Datasets

Several datasets have been used for evaluating word-vector representations. Among them RG [Rubenstein & Goodenough, 1965], WordSim-353 [Finkelstein, et al., 2001], WS-Sim [Agirre, et al., 2009] and MEN [Bruni, Boleda, Baroni, & Tran, 2012]. Each of these datasets consists of word pairs with corresponding similarity scores assigned by human annotators. A model is evaluated by assigning a similarity score to each pair and calculating the correlation (Spearman's ρ) with the human ranking.

However, these datasets suffer from some common shortcomings they have: associations of dissimilar words, low inter-rater agreement over the annotators [Hill, Reichart, & Korhonen, 2016]. In addition, more fundamental problems were pointed out. In some cases the use of rating scales might lead to a variety of annotations biases. In addition, different relations were rated by the same scale and different target-words were rated on the same scale, e.g.: (cat, pet) vs. (winter, season). The mentioned problems were addressed by the method proposed in [Avraham & Goldberg, 2016], however this method requires extensive human annotations.

[4] To calculate PMI matrices we used COMPOSES by [Baroni, Bernardi, & Zamparelli, 2014]

We proposed a simple evaluation scheme that was inspired by [Avraham & Goldberg, 2016], however it does not require extensive human annotations. In addition, our evaluation method can be easily adapted to any language.

In order to evaluate the word-vector representations we picked 32 nouns (hypernyms) with corresponding hyponyms (from two to five for every hypernym). In hypernym-hyponym pairs, the target word (hypernym) with corresponding hyponyms were used to measure the positive pairs of the preferred relations, to measure the negative pairs we used the target words with hyponyms from the different hypernym. To simplify the process, we did not use human annotators to assign similarity scores, the similarity between positive pairs was set to 1 and the similarity between negative pairs was set to 0.

Finally, as a result we calculated Spearman correlation between gold standard 0–1 vector and the vector of cosine similarities calculated in accordance to the tested model.

6. Results

We begin by identifying the best possible settings for stemmer including clustering algorithm and a threshold (Section 6.1). Section 6.2 compares the methods for learning word-vector representations.

6.1. Word Normalization

To find the best performing combination of clustering method and the threshold Θ we ran number of experiments[5]. The preliminary results show that complete-linkage and average-linkage approaches highly outperformed the single-average clustering, so we omit results for the single-average clustering.

According to the evaluation scheme there are *88* positive and *82* negative pairs for hypernym-hyponym relation.

As a baseline approach for comparison we used PMI. The PMI performance score without the stemming was *0.515* for hypernym-hyponym relation. The average-linkage clustering outperformed the complete-linkage clustering. The average-linkage clustering achieves its best results at $\Theta = 1.5$ for hypernym-hyponym relation.

[5] The clustering was performed by fastcluster 1.1.24 [Mullner, 2013]

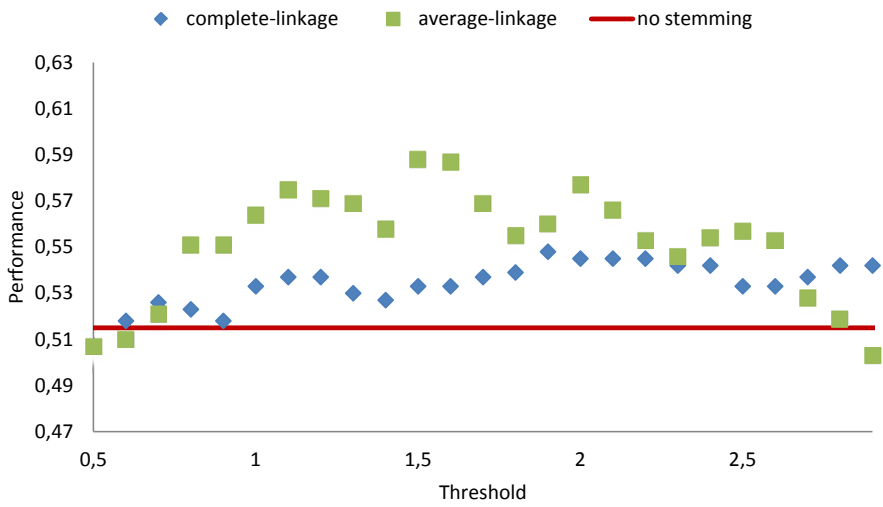


Fig. 1. Evaluation results for various clustering methods and thresholds for hypernym-hyponym relation

6.2. Word-Vector Representations

The performance of the different word embedding methods is in Table 1. Almost in all cases bigger window size leads to better results. Stemming (based on average-linkage with fine-tuned threshold θ) considerably improves word embedding performance.

Table 1. Performance of each method for different settings

Method	win dim	hypernym-hyponym		
		2	5	10
PMI no stemming	—	.517	.510	.528
PMI	—	.585	.588	.611
PMI smoothed	—	.555	.571	.599
SVD	50	.638	.663	.690
	100	.632	.641	**.722**
	500	.612	.662	.691
W2V CBOW	50	.022	−.006	.042
	100	−.003	−.015	.038
	500	−.024	−.033	.011
W2V SGNS	50	.064	.146	.293
	100	.043	.136	.290
	500	.061	.150	.280
GloVe	50	.115	.262	.363
	100	.124	.267	.363
	500	.127	.267	.390

Our findings confirm that SGNS outperforms CBOW on small datasets [Mikolov, Le, & Sutskever, 2013]. In addition, it justifies that Skip-Grams approach works much better on the semantic tasks [Mikolov, Chen, Corrado, & Dean, 2013].

Surprisingly, smoothed variation of PMI (with $\alpha = 0.75$) that was shown to outperform traditional PMI on English Wikipedia corpus [Levy, Goldberg, & Dagan, 2015], lost to traditional PMI when small corpus was used.

To reduce dimensionality we used SVD factorization on PMI matrices after stemming. As expected, SVD factorization outperformed PMI matrices performance in all modes. However, weighted SVD ($d = 0; 0.5$) did not improve the performance further.

Both fails of the count-based methods' enhancements (smoothed PMI and weighted SVD) can be caused by the small size of the dataset.

In addition, traditional count-based methods notably outperformed neural based methods in all settings, which contradicts with the results obtained on big datasets [Baroni, Dinu, & Kruszewski, 2014].

7. Conclusion and Future Work

In this paper we compared the capabilities of traditional count-based methods and neural embeddings methods to learn accurate word-vector representations in small text corpora (on the case of the Buryat Wikipedia). We found that traditional count-based methods outperform neural-based methods when the models are trained on small dataset. We believe that word2vec performance decrease in small corpora was caused by the fact that neural-based models need a lot of training data in order to fit their high number of parameters.

We found that the tweaks (smoothed PMI and weighted SVD) that were found to improve performance on big text corpora [Levy, Goldberg, & Dagan, 2015] did not outperform the traditional PMI and SVD in our case. These fails can be caused by a small size of the dataset. Therefore, future work should carefully explore the influence of the hyperparameters on the quality of the word-vector representations.

We found that language independent stemming approach (with tuned hyperparameters) can considerably improve word embeddings quality.

In addition, we proposed a coarse but easily reproducible word embedding evaluation scheme.

To promote further research, we made our code freely available[6].

8. Acknowledgements

We thank Anton Alexeev, Anna Potapenko, Andrey Kutuzov and Dmitry Ustalov for their assistance and contribution.

[6] https://github.com/vaskonov/burvec

References

1. *Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pacsca, M., Soroa, A..., et al.* (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 19–27.
2. *Altszyler, E., Sigman, M., Ribeiro, S., & Slezak, D. F.* (2016). Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. arXiv preprint arXiv:1610.01520.
3. *Avraham, O., & Goldberg, Y.* (2016). Improving reliability of word similarity evaluation by redesigning annotation task and performance measure. arXiv preprint arXiv:1611.03641.
4. *Badagarov, J., Trosterud, T., & Tyers, F.* (2016). Language Documentation and Language Technologies for Circumpolar Region.
5. *Badmaeva, E., & Francis, T.* (2017). A Dependency Treebank for Buryat. TLT, 1–17.
6. *Baroni, M., Bernardi, R., & Zamparelli, R.* (2014). Frege in space: A program of compositional distributional semantics. Linguistic Issues in Language Technology.
7. *Baroni, M., Dinu, G., & Kruszewski, G.* (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 238–247.
8. *Bruni, E., Boleda, G., Baroni, M., & Tran, N.-K.* (2012). Distributional semantics in technicolor. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, 136–145.
9. *Chen, L.-Y., & Chen, S.-M.* (2007). A new approach for automatic thesaurus construction and query expansion for document retrieval. International Journal of Information and Management Sciences.
10. *Church, K. W., & Hanks, P.* (1990). Word association norms, mutual information, and lexicography. Computational linguistics, 22–29.
11. *Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R.* (1990). Indexing by latent semantic analysis. Journal of the American society for information science.
12. *Di Marco, A., & Navigli, R.* (2013). Clustering and diversifying web search results with graph-based word sense induction. Computational Linguistics, 709–754.
13. *Eckart, C., & Young, G.* (1936). The approximation of one matrix by another of lower rank. Psychometrika, 211–218.
14. *Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., et al.* (2001). Placing search in context: The concept revisited. Proceedings of the 10th international conference on World Wide Web, 406–414.
15. *Fujii, O., & Chimeddorj, A.* (2012). Enhancing Lemmatization for Mongolian and its Application to Statistical Machine Translation. 24th International Conference on Computational Linguistics, 115.
16. *Goldberg, Y.* (2016). A Primer on Neural Network Models for Natural Language Processing. Journal of Artificial Intelligence Research, 345–420.
17. *Harris, Z.* (1954). Distributional structure. Word, 146–162.

18. *Hill, F., Reichart, R., & Korhonen, A.* (2016). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. Computational Linguistics.
19. *Jain, A., Murty, N., & Flynn, P.* (1999). Data clustering: a review. ACM computing surveys, 264–323.
20. *Janhunen, J.* (2006). The Mongolic Languages.
21. *Jivani, A. G.* (2011). A comparative study of stemming algorithms. Int. J. Comp. Tech. Appl, 1930–1938.
22. *Khaltar, B.-O., & Fujii, A.* (2008). A Lemmatization Method for Modern Mongolian and its Application to Information Retrieval. IJCNLP, 1–8.
23. *language, B.* (2017). Buryat language—Wikipedia, The Free Encyclopedia. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Buryat_language
24. *Levy, O., & Goldberg, Y.* (2014). Dependency-Based Word Embeddings. ACL, 302–308.
25. *Levy, O., Goldberg, Y., & Dagan, I.* (2015). Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics, 211–225.
26. *Majumder, P., Mitra, M., Parui, S., Kole, G., Mitra, P., Datta, K..., et al.* (2007). YASS: Yet another suffix stripper. ACM transactions on information systems (TOIS), 18.
27. *Marco, B., Georgiana, D., & German, K.* (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. ACL.
28. *Melamud, O., McClosky, D., Patwardhan, S., & Bansal, M.* (2016). The role of context types and dimensionality in learning word embeddings. arXiv preprint arXiv:1601.00893.
29. *Mikolov, T., Chen, K., Corrado, G., & Dean, J.* (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
30. *Mikolov, T., Le, Q., & Sutskever, I.* (2013). Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.
31. *Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J.* (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 3111--3119.
32. *Mullner, D.* (2013). fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. Journal of Statistical Software, 1–18.
33. *Pennington, J., Socher, R., & Manning, C.* (2014). Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing, 1532–1543.
34. *Porter, M.* (1980). An algorithm for suffix stripping. Program, 130–137.
35. *Rubenstein, H., & Goodenough, J.* (1965). Contextual correlates of synonymy. Communications of the ACM, 627–633.
36. *Rumelhart, D., Hinton, G., & Williams, R.* (1986). Learning representations by back-propagating errors. Nature, 60–88.
37. *Turney, P., & Pantel, P.* (2010). From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research, 141–188.
38. *Utsumi, A.* (2014). A semantic space approach to the computational semantics of noun compounds. Natural Language Engineering, 185–234.
39. *Zhu, Z., Li, M., Chen, L., & Yang, Z.* (2013). Building Comparable Corpora Based on Bilingual LDA Model. ACL, 278–282.

ИНТОНАЦИОННАЯ СТРУКТУРА УСТНОГО РАССКАЗА В КОНТЕКСТЕ НЕЗАВЕРШЕННОСТИ[1]

Коротаев Н. А. (n_korotaev@hotmail.com)
РГГУ, РАНХиГС, Москва, Россия

Ключевые слова: устная речь, русский язык, нарратив, дискурсивная просодия, коммуникативная структура, тема, рема, текстовая незавершенность

HOW INTONATION STRUCTURES SPOKEN NARRATIVES: NON-FINAL PHASE CONTEXTS

Korotaev N. A. (n_korotaev@hotmail.com)
RSUH, RANEPA, Moscow, Russia

Topic—focus articulation in Russian has been mainly studied against isolated utterances. In a categorical sentence, this communicative opposition is reflected in the linear-accentual structure [Paducheva 2015]. For a simple declarative sentence, that would normally mean that the topic (*theme*) comes first and has a rising phrasal accent, while the focus (*rheme*) completes the utterance and is pronounced with a falling accent. At the same time, these formal features do more than just differentiate between topics and foci; they also mark the discourse-semantic category of *phase* [Kodzasov 2009]. In syntactically simple utterances, topics tend to correlate with anticipated continuation, hence *non-final* phase; foci are usually phase-final. As I intend to show in this paper, the non-final phase provides a variety of contexts that challenge the topic—focus distinction. The study is based on the "Stories about presents and skiing"—a collection of prosodically annotated spoken narratives.

In Section 1, I concentrate on issues within a simple clause, where non-final verbal elements often have a fuzzy communicative interpretation. In Section 2, I analyze complex syntactic structures. The data show that non-final clauses may demonstrate both thematic and rhematic properties with regard to their intonation patterns, internal structure and discourse function. Hence, one can claim that some non-final clauses are topics, while

[1] Исследование выполнено при поддержке РНФ, грант № 17-18-01184 («Коммуникативная организация естественного дискурса на звучащих и жестовых языках»).

others are foci. However, a majority of non-final clauses in the analyzed corpus may not be unambiguously attributed to either of these categories. Section 3 provides a pilot study of complex intonation patterns. Only phase distinction being considered, utterances with more than one accentual phrase may follow either (i) the basic adaptation strategy (comprising a non-final rising accent and a final falling accent), or, more often, (ii) a complicated strategy: (a) multiple parallel adaption, (b) consecutive adaptation, or (c) parenthetical strategy.

Keywords: spoken discourse, Russian language, narrative, discourse prosody, communicative structure, topic, focus, phase

Локальная структура устного текста в значительной степени определяется его интонационными характеристиками. При помощи акцентного выделения и ассоциированных с ним тональных движений говорящие выражают широкий спектр дискурсивных значений. Цель настоящего исследования — приблизиться к пониманию того, насколько основные принципы описания русской фразовой просодии, блестяще зафиксированные в таких работах, как [Янко 2008], [Кодзасов 2009] и [Падучева 2015], применимы при сплошной интонационной разметке неподготовленного устного монологического дискурса. Непосредственно в статье решается две задачи. Во-первых, показывается, что в устных рассказах встречается значительное число контекстов, в которых традиционное противопоставление темы и ремы в известной степени ретушируется. Этому посвящены разделы 1 и 2. Во-вторых, демонстрируются некоторые возможности такого подхода к анализу интонационной структуры звучащего текста, при котором базовым противопоставлением выступает не коммуникативная дихотомия темы и ремы, а фазовое различие между завершенностью и незавершенностью (раздел 3).

Материалом исследования послужил корпус «Истории о подарках и катании на лыжах» — просодически аннотированная коллекция устных текстов, порожденных информантами после просмотра серий предъявленных им картинок [Spokencorpora 2013]. Количественные данные приводятся для дополнительно размеченного подкорпуса, составляющего примерно половину объема всего корпуса: в этот подкорпус входит 20 рассказов от 5 информантов, общей продолжительностью звучания 15,5 минуты.

1. Тема vs. рема: моноклаузальные высказывания

Противопоставление темы как предмета, или исходной точки, сообщения и ремы как локуса основного коммуникативного значения высказывания в наиболее выпуклой форме реализуется в синтаксически простых предложениях. Для выражения этих значений используется линейно-акцентная структура высказывания [Падучева 2015]. Рема сообщения, будучи обязательным компонентом этого типа иллокуции, чаще всего произносится с нисходящим

акцентом; тема же — если она есть и выделена интонационно[2] — обычно снабжена тем или иным восходящим акцентом.

Рассмотрим начальный фрагмент одного из рассказов корпуса.

(1) *Pic-RUS_01-f_Ski-R*[3]

1. /Жил-был[3] один молодой \челове<u>к</u>[1].
2. ( ) Этот молодой /человек[3] ( ) любил заниматься \сп<u>о</u>ртом[1].
3. ( ) В то же /время[3] он любил \в<u>ы</u>пить[1].

В примере (1) представлено три отдельных сообщения, реализованных в формате простой клаузы. В каждом из них есть ярко выраженный рематический элемент, кодируемый нисходящим акцентом типа ИК-1: *один молодой человек*, *любил заниматься спортом* и *он любил выпить*. Во всех трех случаях рематическому нисходящему акценту предшествует зеркально противопоставленный ему восходящий акцент типа ИК-3. Тонограмма, соответствующая строкам 2 и 3, показана на рисунке 1. Коммуникативная интерпретация восходящего акцента в этих строках вполне прозрачна: в строке 2 он оформляет именную тему *этот молодой человек*, в строке 3 — адвербиальную тематическую составляющую *в то же время*. Коммуникативный статус начального глагольного элемента *жил-был* уже менее очевиден: глагол стоит в абсолютном начале текста, а потому для его тематической трактовки требуются специальные оговорки.

---

[2] В рамках этого исследования не рассматриваются безударные темы: высказывания, в которых на семантических основаниях можно выделить тему, но это не поддерживается акцентным выделением, трактуются как нерасчлененные. Кроме того, изложение ограничено анализом повествовательных высказываний, содержащих в себе по крайней мере один полноценный ассертивный компонент.

[3] В заголовках примеров из «Историй о подарках и катании на лыжах» используется префикс *PIC-Rus* с дальнейшим указанием индекса текста. Запись отрывков выполнена в рамках системы транскрипции устного русского дискурса, первоначально разработанной для корпуса «Рассказы о сновидениях» [Кибрик, Подлесская (ред.) 2009], с некоторым модификациями. Каждая строка соответствует *элементарной дискурсивной единице* (ЭДЕ); в скобках записываются абсолютные и заполненные паузы; слешами обозначаются движения тона в ударных слогах акцентированных словоформ, стрелками — значимые движения тона вне ударных слогов; пунктуационные знаки в конце строк передают фазово-иллокутивные значения ЭДЕ; цифровые индексы при акцентированных словоформах являются результатом пилотной разметки в терминах интонационных конструкций (ИК) (см. [Брызгунова 1982], [Янко 2001, 2008]).

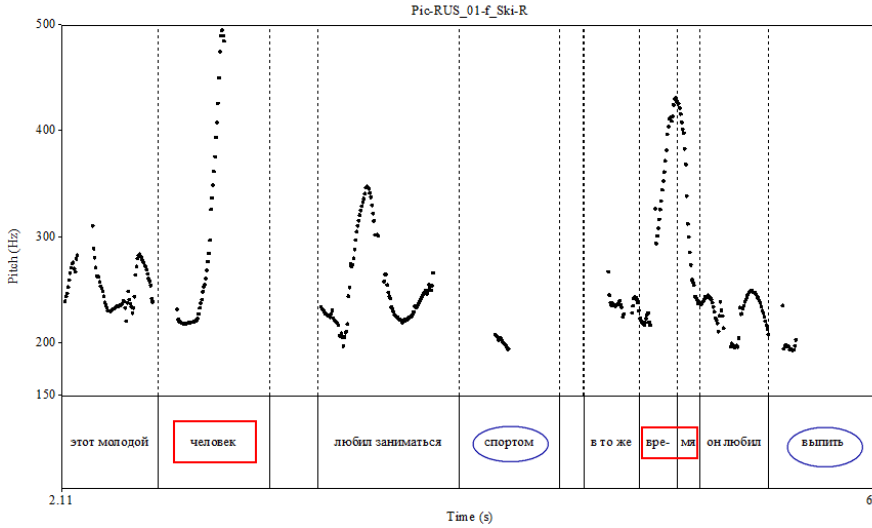


Рис. 1. Тонограмма строк 2–3 примера (1). Овалами выделены слова — акцентносители финального рематического акцента; прямоугольниками — акцентносители восходящего акцента типа ИК-3.

Более убедителен тематический статус начального глагольного элемента *сто́ит она* в строке 33 примера (2): информация о том, что герой истории интересуется ценой автомобиля, содержится в предтексте (строки 26–27)[4].

(2) *Pic-RUS_03-m_Pr-R*

25. ( ) /ну приглянулась ему-у ( ) одна \иномарочка[1].
26. ( ) /Он[3] /решил-л[6] ( ) \узнать[1],
27. /сколько[5(6)] она \стоит[5(1)].
28. ( ) /Подошёл-л[6] к п= || п= || \продавцу-у[1].
29. ( ) /Тот[3] ему \расписал[1],
30. какая она супер-пупер /↓наворо́ченная[6],
31. ( ) (\во-от[1],)
32. ( ) /и /–сказал-л[6],
33. ( ) что-о /стоит[6] она-а ( ) \дорого[2].

Однако столь «надежные» свидетельства тематического статуса начального глагольного элемента простой клаузы крайне редки. В подавляющем большинстве случаев (см. таблицу 1 в конце раздела 2) таких свидетельств нет. В частности, это верно для произносимых с акцентным выделением и интонационно противопоставленных рематическому акценту словоформ *решил*

[4] Коммуникативное членение строки 33 также еще раз демонстрирует известный факт о том, что коммуникативные составляющие необязательно соответствуют составляющим синтаксическим [Янко 2001: 37–38]: последовательность *сто́ит* она не формирует единой синтаксической группы.

(строка 26) и *подошёл* (строка 28) в примере (2)[5]. В целом, именно интерпретация начальных глагольных единиц оказывается одной из главных проблем при сплошной коммуникативной разметке синтаксически простых высказываний.

2. Тема vs рема: поликлаузальные цепочки

В дискурсе не так много изолированных высказываний. При построении «устных предложений» (см. [Chafe 1994: 137–145], [Кибрик 2008]) говорящие часто используют сложные структуры, состоящие более чем из одного дискурсивного шага, в том числе — более чем из одной клаузы. Так, в дополнительно размеченном подкорпусе «Историй о подарках и катании на лыжах» на **49** моноклаузальных сообщений приходится **90** поликлаузальных. В этом случае только в последней по порядку клаузе четко выявляется «классическая» рема с нисходящим движением тона, предшествующие дискурсивные единицы характеризуются *фазовой незавершенностью* [Кодзасов 2009]. Примером сложной поликлаузальной цепочки можно считать последовательность строк 29–33 в примере (2), к анализу которой мы вернемся ниже.

Разберем еще несколько примеров.

(3) *Pic-RUS_05-m_Ski-T*

15. ( ) Из-за того что он ( ) выпил больше чем ( ) /\одну,
16. и-и больше чем две /рюмки[3],
17. ( ) когда он /спускался[3],
18. у /–него[6] ( ) /–одна[6] лыжа /отстегнулась[3],
19. ( ) а /вторая[6] \нет[1].
20. ( ) Перекрутило ему /ногу[3],
21. и сломало её в трёх \–местах[2].

В примере (3) содержится сразу несколько нефинальных клауз: строки 15–16, 17, 18 и 20. В предтексте рассказывается, что герой «хорошенько выпил» и принял опрометчивое решение отправиться на горнолыжную трассу. В связи с этим практически безусловен тематический статус строк 15–16 и 17; этот статус поддерживается и синтаксически: нефинальные отрезки оформлены как препозитивные обстоятельственные придаточные, для которых тематическая функция наиболее характерна[6].

Однако, как и в случае с начальными глагольными элементами в простой клаузе, столь очевидная тематичность нефинальной клаузы — скорее исключение, чем правило. В следующем фрагменте, в котором другая рассказчица

[5] Все три рассмотренные глагольные единицы примера (2) произносится с акцентом типа ИК-6. О взаимозаменяемости ИК-3 и ИК-6 в некоторых контекстах см. [Янко 2008: 67–69].

[6] Еще одним формальным признаком тематичности можно считать некоторые случаи риторических повторов, в которых ранее упомянутая клауза полностью или частично повторяется в качестве фона для описания дальнейших событий [Коротаев 2004].

описывает события, непосредственно предшествующие событиям из примера (3), имеется цепочка из трех нефинальных и одной финальной клаузы:

(4) *Pic-RUS_02-f_Ski-R*

4. ( ) (ə) (ʔ) ( ) (ɯ) Во время этого он встретил своих /друзе-ей[3],
5. которые решили-и (ə) отметить /встрéчу-у[3],
6. (ʔ) во время \отмечания он ( ) (ɯ) /перебрá-ал[3],
7. и-и (ə) в пьяном /состоянии[3] решил \снóва[1] покататься на лыжах.

В [Янко 2008: 141–148] для серий неконечных клауз, произносимых с акцентами типа ИК-3, используется термин *множественные темы*: события, описываемые в этих клаузах, трактуются как «сюжетный фон для какого-то более важного события» (с. 142). Нам, однако, представляется, что строки 4–6 примера (4), а также строки 18 и 20 примера (3) и строки 29–30 примера (2), можно интерпретировать и по-другому: а именно, как ремы с наложенным значением незавершенности. Приведем несколько аргументов в пользу такой трактовки.

А. С семантической точки зрения доказать меньшую значимость событий, описываемых в нефинальных клаузах, не всегда легко. Так, в примере (4) встреча героя с друзьями, превращение встречи в застолье, чрезмерное увлечение героем крепкими напитками и его решение вернуться на лыжную трассу — события, примерно равнозначные для развития истории. Еще очевиднее семантическое равноправие нефинальной и финальной клауз в паре строк 18–19 из примера (3), где, по сути, сложным образом описывается единая ситуация драматической рассинхронизации лыж при спуске.

Б. Нефинальные клаузы могут обладать самостоятельной внутренней расчлененностью, причем в некоторых случаях эта расчлененность дублирует коммуникативную структуру финальной клаузы. В строках 18–19 примера (3) реализовано двойное контрастивное противопоставление: *одна лыжа* (внутренняя тема строки 18) и случившееся с ней противопоставляется *второй* (внутренняя тема строки 19) и (не) случившемуся с этой лыжей. В строке 6 примера (4) роль внутренней темы играет адвербиальная группа *во время отмечания*, ссылающаяся на упомянутый в предыдущей строке характер встречи; точно так же в финальной строке 7 внутренняя тема *в пьяном состоянии* непосредственно отсылает к глаголу *перебрал* из строки 6. По сути, тут представлен стандартный механизм поддержания дискурсивной связности: рема предыдущего высказывания переходит в тему следующего[7]. Заметим также, что внутренние темы могут интонационно адаптироваться как к конечному движению тона в рамках нефинальной клаузы (*во время отмечания* в примере (4)), так и к заключительному рематическому акценту в финальной ЭДЕ (*одна лыжа* в примере (3)). Подробнее о разных способах адаптации см. раздел 3.

[7] Отметим, что при подсчетах, приведенных в таблице 1, в качестве отдельных единиц учитывались и внутренние тематические элементы в составе нефинальных клауз, и нефинальные клаузы целиком. Например, для строки 6 примера (4) отдельно учитывались единицы *во время отмечания* и *во время отмечания он перебрал*.

В. В нефинальной клаузе может выражаться самостоятельное иллокутивное значение — см. строку 28 следующего примера, в которой незавершенность накладывается на иллокуцию директива:

(5) *Pic-RUS_05-m_Pr-T*

27. ( ) *продавец* *говорит*[1]:
28. «/–*Покупайте*[6]¡,
29. вот это самый лучший /автомобиль[3],
30. из всех что \есть[1].»

При отсутствии незавершенности директив скорее произносился бы с нисходящим акцентом. Примеры типа (5) *per se* в настоящем исследовании не рассматриваются, однако в них наглядно иллюстрируется принципиальная возможность совмещения иллокутивного значения со значением незавершенности.

Г. При так называемой аналитической стратегии выражения незавершенности коммуникативное (рема) и фазовое (незавершенность) значения кодируются отдельными акцентами [Янко 2008: 128–141]. Так, в строке 11 примера (6) нисходящий акцент типа ИК-1 на *ещё раз* маркирует рему, а восходящий акцент типа ИК-3 на *лыжах* отвечает за незавершенность:

(6) *Pic-RUS_04-m_Ski-T*

10. ( ) У него за-а= ( ) =кружилась /голова[3],

11. и он решил ещё \раз[1] прокатиться на /лыжах[3],

12. то= || но только с /–большо-ой[5(6)] –\горки[5(2)].

При совместном выражении коммуникативного и фазового значений единственный акцент в данном контексте приходился бы на *ещё раз*, а глагольная группа *прокатиться на лыжах* произносилась бы безударно:

(6') и он решил ещё /раз прокатиться на лыжах,

При этом вопрос о том, утрачивает ли при таком произнесении группа *ещё раз* рематические свойства, имеющиеся у нее в строке 11 примера (6), остается открытым.

Итак, «огульная» тематическая трактовка нефинальных клауз, произносимых с восходящей интонацией типа ИК-3 / ИК-6, может быть оспорена при помощи ряда семантических, дискурсивных и формальных аргументов. Впрочем, верно и обратное: в большинстве случаев (см. таблицу 1 в конце текущего раздела) отсутствуют и веские основания признавать такие клаузы ремами. Более того, формальные тематические и рематические признаки могут накладываться друг на друга. Например, в строке 25 следующего фрагмента, с одной стороны, нефинальная клауза синтаксически оформлена как препозитивное временно́е придаточное, с другой стороны, использована аналитическая стратегия выражения незавершенности:

(7) *Pears22C8*

25. \первый[1] раз когда он /слезает[6],
(ц)
26. он \докладывает[1] туда до /верха[3],
27. и у него остаётся /третья[3] /–пустая[6] /\корзина[2].

Помимо ИК-3 / ИК-6 существуют и другие способы интонационного кодирования незавершенности. При одном из них говорящие, перечисляя события рассказа, произносят нефинальные клаузы с нисходяще-восходящим акцентом типа ИК-4 — см. строки 16 и 17 примера (8):

(8) *Pic-RUS_06-f_Pr-T*

14. ( ) (ə) Он \спросил-л[1],
15. у другого /\мужчины-ы[1],
16. ( ) (ш) сколько она \↑стоит[4],
17. ( ) {смех} ( ) /тот[3] сказал ему что-о {смех} очень \/↑много[4],
18. ( ) {смех} ( ) \и-и (ʔ) ( ) /тогда[3] ( ) этот /дядечка[3] решил не \покупать[1] \ машину.

Использование ИК-4 характерно при педантичном «рассказе по порядку» [Янко 2008: 167–170]. В отличие от дефолтного способа выражения незавершенности при помощи ИК-3, такое интонационное оформление не типично для тематических элементов, а потому — с некоторой долей условности — может считаться просодическим признаком ремы. И все же «не типично» не означает «невозможно»: в корпусе изредка встречаются внутриклаузальные именные темы, произносимые с ИК-4, — см. *один дядечка* в строке 1 следующего примера:

(9) *Pic-RUS_09-f_Pr-T*

1. ( ) В один прекрасный солнечный /день[3] ( ) (ə) ( ) один –↑дядечка[4] ( ) (ə) /вспомнил[3],
2. что-о у /–жены[6] →его ( ) день \рождения[1].

Мы рассмотрели ряд контекстов, в которых определение тема-рематического статуса интонационно выделенного отрезка речи существенно затруднено. Отсюда не следует, что это традиционное коммуникативное противопоставление вовсе не применимо к анализу устного неподготовленного дискурса. Однако при сплошной разметке устных нарративов в терминах тем и рем слишком часто приходится опираться на нетривиальные теоретические решения — а значит, результаты такой разметки могут оказаться недостаточно надежными и плохо воспроизводимыми.

В таблице 1 представлены количественные данные о рассмотренных явлениях в размеченном подкорпусе — совместно для моно- и поликлаузальных структур.

---

[8] Пример (7) взят из другого устного корпуса — «Рассказов и разговорах о грушах», литера С в заголовке примера обозначает роль говорящего в записи; подробнее см. www.multidiscourse.ru.

Таблица 1. Типы нефинальных единиц в дополнительно размеченном подкорпусе

Тип единицы		Кол-во	%
Тематические ИГ		100	21,1
Адвербиальные темы		51	10,8
Нефинальные глагольные элементы		121	25,6
- в т. ч. с «надежными» тематическими свойствами	*17*		
Нефинальные клаузы		176	37,2
- в т. ч. с формальными признаками темы	*20*		
- в т. ч. с формальными признаками ремы:	*46*[9]		
(а) внутренняя расчлененность	25		
(б) интонационное оформление (ИК-4 и др.)	25		
(в) аналитическое выражение незавершенности	3		
Прочее		25	5,3
Итого		**473**	**100**

3. Case-study: стратегии адаптации при незавершенности

В ряде примеров, рассмотренных в предыдущих двух разделах, затруднена коммуникативная интерпретация некоторых отрезков в терминах тем и рем. Однако *фазовые* значения этих отрезков кажутся более очевидными: характер тоновой кривой вполне однозначно указывает на завершенность или незавершенность[10].

В связи с этим разумно поставить следующий исследовательский вопрос: можно ли выделить повторяющиеся паттерны интонационно расчлененных устных структур, игнорируя коммуникативные категории и опираясь только на противопоставление завершенности и незавершенности? Ниже приведены предварительные результаты, полученные в рамках такого подхода.

В дополнительно размеченном подкорпусе были проанализированы все цепочки, состоящие из ненулевого количества отрезков с интонационно выраженной незавершенностью (далее — *нефинальные отрезки*) и одного отрезка с интонационно выраженной завершенностью, «разрешающей» предшествующую незавершенность (далее — *финальный отрезок*). Не рассматривались последовательности, не имеющие явно выраженной конечной завершенности,

[9] Одна клауза может обладать сразу несколькими формальными признаками ремы, поэтому сумма чисел для пунктов (а)–(в) превышает общее количество в этой строке.

[10] В действительности, это некоторое упрощение. В частности, в речи рассказчиков регулярно встречаются так называемые «нефинальные падения», сигнализирующие о неполной завершенности цепочки [Кибрик 2008]. В настоящем исследовании эта деталь игнорируется и большинство нефинальных падений считаются признаками завершенности.

а также цепочки, не завершающие иллокуции сообщения[11]. Задачей пилотного исследования было выявить основные стратегии интонационной адаптации нефинальных отрезков к финальному, а также обозначить частные закономерности, касающиеся распределения этих стратегий в материале подкорпуса. Метод анализа частично наследует принципам классификации интонационных стратегий оформления сложноподчиненных и цитационных конструкций, изложенным в [Коротаев, Подлесская 2008] и [Подлесская 2017].

Образцы исследованных цепочек в большом количестве представлены в уже рассмотренных выше примерах. В частности, в примере (9) имеется последовательность из четырех нефинальных отрезков: *в один прекрасный солнечный день*, *один дядечка*, *вспомнил* и *что у жены его*; а строках 20–21 примера (3) — один нефинальный отрезок *перекрутило ему ногу*. Во втором случае реализуется ***базовый принцип адаптации***, многократно описанный в литературе и уже упомянутый выше: нефинальный отрезок произносится с восходящим акцентом, адаптирующимся к нисходящему акценту в финальном отрезке (*и сломало её в трёх местах*). Аналогичным образом, но уже в пределах моноклаузальной конструкции, в каждой из трех строк примера (1) единственный нефинальный отрезок (жил-был; этот молодой человек; в то же время) базово адаптируется к финальному (один молодой человек; любил заниматься спортом; он любил выпить).

В свою очередь, в примере (9) реализовано одно из возможных усложнений базового принципа — ***множественная параллельная адаптация***: каждый из нефинальных отрезков произносится с тем или иным видом восходящего акцента (ИК-3, ИК-4, ИК-6) — вне зависимости от своего синтаксического статуса. В итоге, с точки зрения интонационного оформления, все нефинальные отрезки равноправно адаптируются к финальному. То же верно и для строк 15–19 примера (3), в которых содержится семь нефинальных отрезков. Как уже обсуждалось в разделе 2, большинство этих отрезков проявляют тематические свойства, но глагол *отстегнулась* скорее завершает рематическую составляющую. Для анализа в терминах фазовых значений, однако, это различие не принципиально.

Противоположная стратегия оформления сложных цепочек — это ***последовательная адаптация***. В этом случае только в одном нефинальном отрезке движение тона непосредственно адаптируется к движению тона в финальном отрезке, тогда как предшествующий отрезок адаптируется к ближайшему правому контексту. Фактически это означает, что этот отрезок произносится с тем или иным видом адаптивного нефинального падения — см. *купил* в строке 20 примера (10). Последовательная адаптация может реализовываться на протяжении всей цепочки, но чаще эта стратегия комбинируется с множественной параллельной адаптацией:

[11] Таким образом, финальный отрезок в данном случае всегда является либо ремой повествовательного высказывания, либо ее частью.

(10) *Pic-RUS_06-f_Pr-T*

19. ( ) (ə) ( ) /Пошёл[3],
20. \купил[1] ( ) \↑маленькую[4] машинку,
21. ( ) /подарил[3] её \жене[1].

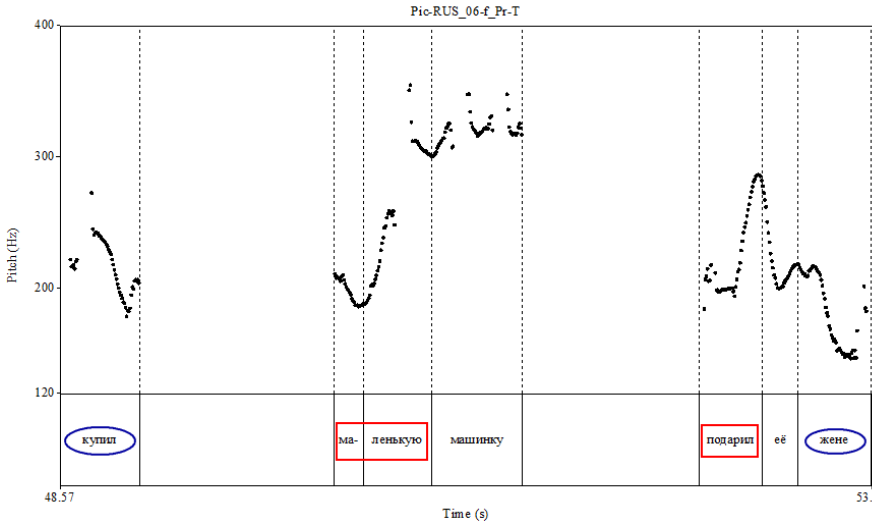


Рис. 2. Тонограмма строк 20–21 примера (10). Овалами выделены слова — акцентносители нисходящего акцента; прямоугольниками — акцентносители восходящего акцента

В примере (10) нефинальные отрезки *пошёл*, *маленькую машинку* и *подарил* интонационно адаптируются к финальному отрезку *её жене* согласно стратегии множественной параллельной адаптации. При этом глагольная форма *купил* снабжена нефинальным нисходящим акцентом, адаптирующимся к восходящему акценту типа ИК-4 в ИГ *маленькую машинку*: таким образом, на этом участке использована стратегия последовательной адаптации. Тонограмма, соответствующая строкам 20–21 примера (10) приведена на рисунке 2. Похожим образом устроен и пример (4). В целом он следует стратегии множественной параллельной адаптации, но в строке 6 представлен случай последовательной адаптации: нисходящий акцент на словосочетании *во время отмечания* адаптирован к ближайшему восходящему акценту на *он перебрал*, который, в свою очередь, противопоставлен финальному падению в строке 7.

Наконец, еще одно важное усложнение базового принципа адаптации — это ***вставка***. При вставке вслед за одним из нефинальных отрезков говорящий произносит отдельное высказывание — обычно с нисходящим акцентом. Особенность этого вставленного высказывания заключается в том, что оно не может считаться заключительным к имеющейся незавершенности; финальный отрезок данной последовательности будет произнесен позже, после выхода

из вставки. Как и в случае с последовательной адаптацией, вставка чаще реализуется совместно с множественной параллельной адаптацией. В примере (11) оба нефинальных отрезка в строке 18 и единственный нефинальный отрезок в строке 21 произносятся с восходящими акцентами, параллельно адаптированными к финальному нисходящему акценту типа ИК-2 на *несчастный*; а в строках 19–20 реализуется вставка:

(11) *Pic-RUS_01-f_Ski-T*

18. ( ) На /его[3] наложили /\/гипс[4],
19. ( ) (туда и ↓–сюда[1],
20. ( ) на ногу и \на[1] руку то есть,)
21. ( ) /теперь[3] он ходит ↑\несч-частный[2].

Во вставленных строках, каждая из которых произносится с нисходящим акцентом, говорящая временно приостанавливает развитие основной линии изложения, уточняет информацию, содержащуюся в строке 18, но не «ставит точку» в поликлаузальной цепочке. Продолжение — и завершение — цепочки происходит уже после вставки, в строке 21.

Из **146** проанализированных цепочек базовому принципу адаптации следуют **39**, в остальных **107** цепочках имеется хотя бы одно из описанных выше усложнений базового принципа. В строках 29–30 примера (2), повторяемых ниже под номером (12), реализованы все три усложняющие фигуры: последовательная адаптация на участке *ему расписал какая она супер-пупер навороченная*, вставка в строке 31 и множественная параллельная адаптация в остальных нефинальных отрезках.

(12) *Pic-RUS_03-m_Pr-R*

29. ( ) /Тот[3] ему \расписал[1],
30. какая она супер-пупер /↓наворочennая[6],
31. ( ) (\во-от[1],)
32. ( ) /и /–сказал-л[6],
33. ( ) что-о /стоит[6] она-а ( ) \дорого[2].

Стратегии множественной параллельной и последовательной адаптации разграничиваются на основании сугубо формальных признаков. Однако между ними имеются и заслуживающие отдельного упоминания содержательные различия. Эти различия наблюдаются, в частности, в примере (12). Можно заметить, что отрезки *ему расписал* и *какая она супер-пупер навороченная*, входящие в фигуру с последовательной адаптацией, формируют замкнутую синтаксическую составляющую — глагольную группу с сентенциальным зависимым. Внутренняя связь между отрезками этой составляющей теснее, чем связь любого из отрезков по отдельности с каким-либо отрезком из левого или правого контекста. В то же время ни для какой смежной пары нефинальных отрезков примера (12), следующих стратегии множественной параллельной адаптации, обнаружить такую замкнутую составляющую нельзя.

Общие данные о наличии / отсутствии замкнутой содержательной связи между нефинальными отрезками на участках с множественной параллельной и последовательной адаптацией представлены на рисунке 3. Как видно, в целом распределение следует описанному выше стандарту: связь такого рода наблюдается более чем в 75 % фрагментов со стратегией последовательной адаптации, но лишь в 15 % случаев стратегии множественной параллельной адаптации. Данное различие обладает статистической значимостью ($p < 0.00001$; точный критерий Фишера).

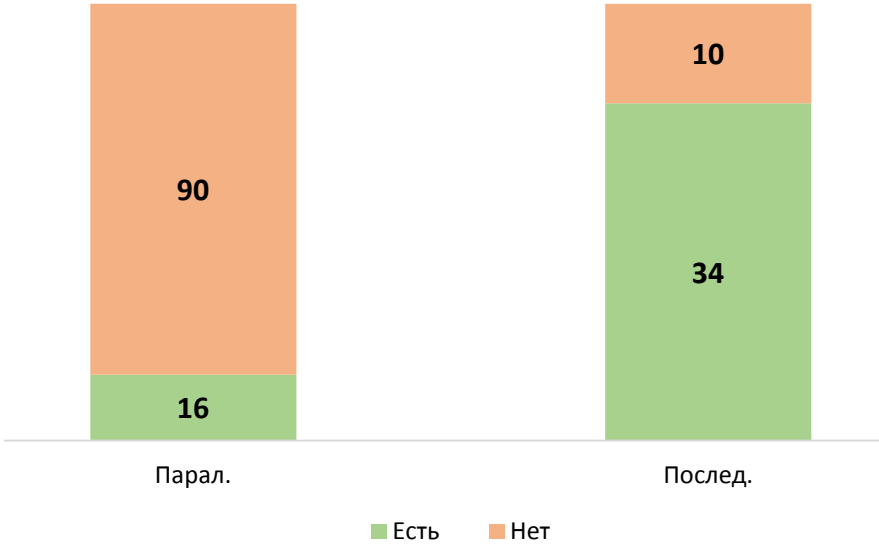


Рис. 3. Замкнутая связь нефинальных отрезков при стратегиях множественной параллельной и последовательной адаптации

4. Заключение

При попытке сплошной коммуникативно-просодической разметки корпуса неподготовленных устных рассказов выявляется существенное количество контекстов, в которых традиционное разграничение темы и ремы представляет нетривиальную теоретическую задачу. В то же время анализ устного текста в терминах фазового противопоставления завершенности / незавершенности позволяет выявить набор основных стратегий интонационной адаптации, используемых рассказчиками при построении сложных высказываний. Так, при стратегии последовательной адаптации говорящие обычно подчеркивают локальную дискурсивную иерархию посредством скобочной акцентной структуры. При стратегии множественной параллельной адаптации, напротив, этого чаще всего не происходит. Исключения тут немногочислены и могут быть подвергнуты дальнейшей каталогизации при увеличении объема исследованного материала.

Литература

1. *Bryzgunova E. A.* (1982), Intonation [Intonacija], Russian Grammar [Russkaja grammatika], Vol. 1, Nauka, Moscow, pp. 98–118.
2. *Chafe W.* (1994), Discourse, consciousness, and time, University of Chicago Press, Chicago.
3. *Kibrik A. A.* (2008), Is sentence viable in spoken discourse? [Est' li predloženie v ustnoj reči?], Phonetics and non-phonetics. A 70th birthday Festshcrift for Sandro V. Kodzasov [Fonetika i nefonetika. K 70-letiju Sandro. V. Kodzasova], LRC, Moscow, pp. 104–115.
4. *Kibrik A. A., Podlesskaja V. I.* (eds.) (2009), Night Dream Stories: A corpus study of spoken Russian discourse [Rasskazy o snovidenijax: Korpusnoe issledovanie ustnogo russkogo diskursa], LRC, Moscow.
5. *Kodzasov S. V.* (2009), Studies in Russian Prosody [Issledovanija v oblasti russkoj prosodii], LRC, Moscow.
6. *Korotaev N. A.* (2004), Situational anaphora in Russian spoken narratives [Situacionnaja anafora v ustnom russkom narrative], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2004" [Komp'juternaja lingvistika i intellektual'nye texnologii: Trudy Meždunarodnoj konferencii "Dialog-2004"], Issue 3 (10), Moscow, pp. 334–345.
7. *Korotaev N. A., Podlesskaya V. I.* (2008), Prosody of clause-combining in Russian: A corpus-based study [Frazovaja akcentuacija v složnyx predloženijax s postpozitivnym pridatočnym v russkom jazyke: opyt ispol'zovanija ustnogo korpusa s prosodičeskoj razmetkoj], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2008" [Komp'juternaja lingvistika i intellektual'nye texnologii: Trudy Meždunarodnoj konferencii "Dialog-2008"], Issue 7 (14), Moscow, pp. 234–240.
8. *Paducheva E. V.* (2015), Communicative structure of the sentence [Kommunikativnaja struktura predloženija], Materials for a corpus-based Russian grammar [Materialy dlja proekta korpusnogo opisanija russkoj grammatiki], Moscow. Online: http://rusgram.ru/.
9. *Podlesskaya V. I.* (2017), "Ja skazhu tebe s poslednej prjamotoj": Direct and indirect speech viewed through the prism of prosodically annotated corpus data ["Ja skažu tebe s poslednej prjamotoj": prjamaja i kosvennaja reč' po dannym korpusa s prosodičeskoj razmetkoj], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2017" [Komp'juternaja lingvistika i intellektual'nye texnologii: Trudy Meždunarodnoj konferencii "Dialog-2017"], Issue 16 (23), Vol. 2, Moscow, pp. 355–371.
10. *Spokencorpora* (2013) Prosodically Annotated Corpus of Spoken Russian (PrACSRuss). Pilot version. Online: http://spokencorpora.ru.
11. *Yanko T. E.* (2001), Communicative strategies of Russian speech [Kommunikativnye strategii russkoj reči], LRC, Moscow.
12. *Yanko T. E.* (2008), Intonation strategies of Russian speech from a contrastive perspective [Intonacionnye strategii russkoj reči v sopostavitel'nom aspekte], LRC, Moscow.

FRAMES REVISITED: AUTOMATIC EXTRACTION OF SEMANTIC PATTERNS FROM A NATURAL TEXT[1]

Kotov A. A. (kotov_aa@nrcki.ru),
Zaidelman L. Y. (zaydelman_ly@nrcki.ru),
Arinkin N. A. (arinkin_na@nrcki.ru)

Kurchatov Institute; Russian State University for the Humanities; Moscow, Russia

Zinina A. A. (zinina_aa@nrcki.ru)

Kurchatov Institute, Moscow, Russia

Filatov A. A. (alexander.filatov@hp.com)

HP Inc., Moscow, Russia

Our project aims to design a syntactic parser, which constructs a semantic representation in a frame format: a clause is represented as a table of valencies, filled in with semantic markers. This representation is compared to a list of scripts—used to disambiguate and classify the semantic representation as well as to select an appropriate reaction for a companion robot F-2.

Key words: semantic frames, syntax parser, semantic analysis, sentiment analysis

[1] The development of semantic patterns and scripts is supported by RFBR grant 16-29-09601; the analysis and classification of text threats with the parser is supported by RSF grant 17-78-30029.

ВОЗВРАЩЕНИЕ К ФРЕЙМАМ: АВТОМАТИЧЕСКОЕ ВЫДЕЛЕНИЕ СЕМАНТИЧЕСКИХ ПАТТЕРНОВ ИЗ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЫКЕ

Котов А. А. (kotov_aa@nrcki.ru),
Зайдельман Л. Я. (zaydelman_ly@nrcki.ru),
Аринкин Н. А. (arinkin_na@nrcki.ru)
Курчатовский институт; Российский государственный гуманитарный университет; Москва, Россия

Зинина А. А. (zinina_aa@nrcki.ru)
Курчатовский институт, Москва, Россия

Филатов А. А. (alexander.filatov@hp.com)
HP Inc., Москва, Россия

1. Introduction

The notion of *frame* [Minsky, 1988] has been a key concept for semantics for several decades, until it was recently replaced by a network approach where information is stored in a trained classifier and is not readily available for linguistic analysis. We developed an automatic system for the analysis of natural language text, designed to invoke emotional reactions (gestures, facial expression, speech replies) from the meaning of incoming texts. The system can animate a companion robot F-2 or can store the extracted meaning with a suggested emotional reaction to a database. In this work we describe the processing of the incoming text up to its meaning which is stored to the database. We return to the classical form of semantic representation—*frame*, and describe semantics as a set of semantic features (markers) divided up into a set of valencies (*agens, patient,* etc.). To construct the semantic representation we develop a parser, implementing morphological, syntactic and semantic processing of an incoming text as suggested by theoretical linguistic models, e.g. [Melčuk, 1999]. The parser is written in C#, the grammar of Russian is described in syntXML language and the dictionary is stored in an SQL database. On each step of processing, the parser may save the intermediate results of analysis to an SQL database or transfer them to the next software component—Fig. 1. In this work we present a general architecture of the parser and the methods to analyze the semantic representations, extracted during a text analysis.

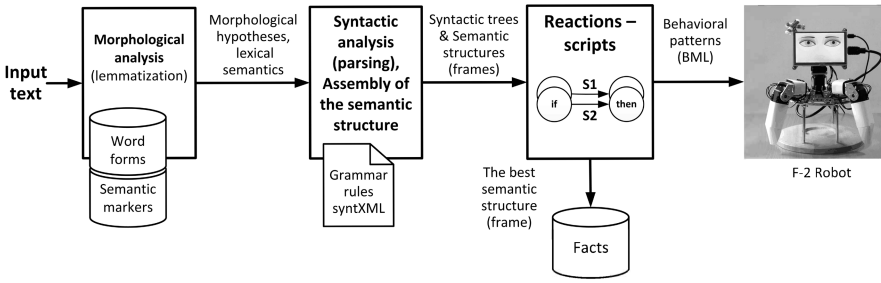


Fig. 1. The architecture of the text analysis

2. The architecture of the parser

Scripts (the models of emotional reactions) are designed to detect emotional (input) patterns in incoming texts [Kotov, in press]. The representations, to be searched by scripts, are quite shallow: '*X* is a tasty food' or 'I was hit by someone' are sufficient patterns to suggest an emotional reaction to the robot. To detect the patterns, parser should store the main taxonomic markers for natural words—suggest that *John* is 'a person' and *strawberry* is 'a food', as well as the emotional markers—*idiot* is 'an inadequate person'. The parser should also allocate the markers to specific valencies in order to distinguish the antagonist and protagonist roles in an emotional event. Emotions (unlike formal logical notation) are sensitive to the syntactic structure of a sentence: as suggested in a classic work by Blakar, *The police took in the demonstrators* sounds more violent and emotional than *The demonstrators were taken in by the police* [Blakar, 1979]. This implies that the changes of diathesis should be revealed in the required semantic representation in a form of semantic/syntactic roles—valencies. Therefore, the task of the parser is to construct a semantic representation, where semantic markers are applied to a set of valencies. This exactly correspond to the classic representation of *frame*. Table 1 shows the semantic representation of an utterance *Linguists have noticed psychologists at the conference* (*Lingvisty otmetili psihologov na konferencii*). Such semantic representations can be quite shallow, sufficient only to detect the input patterns for the activation of scripts. In the architecture of the parser, scripts play the major role—they select the best tree with the precision, sufficient to extract the input patterns. The distinction between trees, equally weighted by the emotions, is ignored.

Table 1. Semantic representation (frame) of the sentence *Linguists have noticed psychologists at the conference*

Valency	*p*	*ag*	*cont*	*loc*
Semantic markers	think pay-attention	somebody profession	somebody profession	abstract-entity container abstract-container

2.1. Morphological processing: lemmatization and dictionary

For a given wordform, the lemmatizer should extract (a) a required set of semantic markers and (b) a set of grammemes to construct a syntactic tree and further allocate the valency for the specific word. The morphological dictionary is based on the OpenCorpora project [Granovsky, Bocharov, Bichineva, 2010] and stores all wordforms and grammemes for 98,000 lemmas. To each incoming segment the lemmatizer assigns a number of morphological hypotheses. In case no hypothesis is assigned, a segment is consecutively processed by (a) a user dictionary, where one can store a noun with an inflectional class and (b) a guesser, which assigns 5 morphological hypotheses, based on a trained Keras neural network (for standard words this offers 96,7% recall). Words from a user dictionary and guesser do not get any semantic representation and are considered as semantically void. 30,000 words in the main dictionary are manually annotated by semantic markers (from 1 to 18 markers per word). We rely on semantic markers from two main groups: taxonomy markers and emotional markers.

(a) **Taxonomy markers** are used to classify the word and the whole predication. We use markers like 'somebody', '[named after] profession', 'move', 'move body', here we rely on the annotation suggested in the Russian semantic dictionary [Shvedova, 1998]. Unlike in traditional ontologies, a word may keep semantic markers from different classes: *bank* has the markers for 'organization', 'building' and 'abstract container'. This polysemy allows us to simulate *situational effects* [Yeh, Barsalou, 2006], where a word meaning is shifted depending on the situation or the emotion to be invoked by semantics—top-down emotional processing [Clore, Ortony, 2000]. In our case, different scripts may detect different markers in the lexical semantics, addressing different word meanings.

(b) **Emotional markers**, on one side, are markers, explicitly expressing a position in an emotional situation and an emotional evaluation like words *fool*, *great* or *useless*. These words are annotated by specific script markers: 'agent/antagonist in a situation of inadequate actions' (INADEQ script), 'protagonist in a situation of superiority' (SUPER), 'antagonist/excessive for the actions' (FREEDOM) or 'protagonist/someone, incorrectly ignored' (UNNEED). On the other side, we annotate ambiguous markers, that may contribute to emotions—*emotionally sensitive markers*, like 'intensity'—relevant in the emotional phrases like *Why do you push me?* [Glovinskaya, 2004].

Emotional markers have attributes, indicating their position in the focal/background semantics of the corresponding word. For example, the word *fool* has the markers 'inadequate' in its focal zone, and 'somebody' ('human') in its background zone, while the word *man* has the marker 'somebody' in its focal zone.

Words may have different meanings (homonymic and polysemic), that is represented by the number of subsets of semantic markers—one subset for each meaning. Different input patterns of scripts may address different markers in the semantics of a word. A script will select the meaning that fits best its input pattern.

2.2. Syntactic processing: syntXML

The task of the syntactic processor is to construct a syntactic representation where each actant gets its valency. The parser implements a left-to-right processing and combines rule-based approach to construct syntactic links with the trained classifier approach to evaluate the syntactic structures. Syntactic rules are designed on syntXML language [Kotov, Zinina, Filatov, 2015]. A rule is defined as a possible reduction, where the right-hand side (one or many segments *a, b, … n*) can be reduced to the left-hand head ***h*** (1). Head ***h*** can also be a member of the right-hand side and subordinate other segments (2).

$$\boldsymbol{h} \rightarrow <a, b, \ldots n> \quad (1)$$

$$\boldsymbol{h} \rightarrow <a, b, \boldsymbol{h}, \ldots n> \text{ or } <a, b, \boldsymbol{h}^{\text{head}}, \ldots n> \quad (2)$$

Right-hand side of a rule may contain one or many segments *a, b, … n*, segments except *n* can be optional. This allows us to describe immediate constituent rules, where ***h*** is not a member of the right-hand side rule section, as well as dependency rules—both with any number of subordinate segments, so the grammar is not limited to binary relations. A rule can (a) verify if a segment matches a specific lexeme, (b) check the presence of a particular grammeme within a segment, (c) check grammatical agreement between rule segments, (d) set or rewrite grammemes for the rule head ***h***, (e) copy grammemes/markers to the rule head ***h***. This combination is represented in an XML form—as a syntXML rule. The grammar for Russian uses 550 rules and structurally corresponds to the System of Syntactic Groups by Gladkiy [Gladkiy, 1985]. In particular—conjunctions are defined as immediate constituents with a virtual head, as the features of the whole construction cannot be fully represented by any segment, whereas other grammar structures are represented as dependency trees.

Each incoming segment is added to a stack and the parser tries to reduce the stack head with a grammar rule, upon the reduction the subordinate segments are excluded from the stack and are linked to the rule head. Morphological and syntactic ambiguities are accounted: a separate stack instance is constructed for each morphological hypothesis and for each ambiguity in the application of syntactic rules. At the same time, the maximum number of stacks *m* is limited, currently by $m = 512$. Each rule application is evaluated, taking into account the corpus data (basing on a converted SynTagRus corpus): co-occurrence of (a) particular words within this rule, (b) word2vec vectors of the words within the rule, (c) sets of grammemes of the words within the rule. In other words, a syntactic link is more probable, if it exists in the corpus between: (a) the same words, (b) similar words, where similarity is calculated by word2vec distance, or (c) words with the same sets of grammemes. An aggregated score is calculated for each stack, after that *m* best stacks are preserved for further analysis. Stacks with lower scores (stacks with less probable syntactic links and stacks with few syntactic links) are discarded. In case several trees are generated by the end of the sentence (the possible number is from 0 to *m*), the best tree is determined and selected depending on its comparison with scripts, as explained in 2.3.

The semantic structure for a tree is constructed in the following way. For each predicate (finite verb, predicative, etc.) a valency ***p*** is assigned. Once a rule binds

a predicate with an actant (usually, a noun phrase), it may assign a valency to the noun phrase. In our project we rely on the list of valencies, suggested by [Fillmore, 1968], and use a list of 22 valencies: *ag* (agent), *pat* (patient), *instr* (instrument) etc. We have marked subcategorization frames for 13,000 Russian verbs, in part, relying on data from FrameBank project [Lyashevskaya, 2010]. A verb gets a number of syntactic markers, where each marker allows the verb to trigger a specific rule to assign an actant to a particular valency. For example, the subcategorization frame of the verb *zvenet'* ('to ring'—*звенеть*) has 8 such valencies, as shown in Table 2.

Table 2. Subcategorization frame of the verb ***zvenet'*** ('to ring'—***звенеть***)

Actant: case and preposition	NP, nominative, no prep.	NP, instrumental, no prep.	NP, accusative, prep. *v* (in)	NP, genitive, prep. *ot* (from)	NP, genitive, prep. *u* (at)	NP, accusative, prep. *o* (about)	Adverb or adverbial NP	Adverb or adverbial NP
Valency	***ag*** (agent)	***instr*** (instrument)	***targ*** (target)	***caus*** (cause)	***pos*** (possessor)	***ca*** (counterparty)	***t*** (time)	***loc*** (location)

Thus, the syntactic structure of the sentence *U menya zvenit v ushah* ('It rings in my ears'—*у меня звенит в ушах*) will assign the predicate *zvenit* ('rings') to ***p*** valency, NP *u menya* (~'at me') to ***pos*** valency (this triggers 1-y-Sgen-pos marker of the predicate) and NP *v ushah* ('in ears') to ***loc*** valency (this triggers 1-loc marker).

Following the syntactic analysis, the semantic representation is constructed based on the lexical semantics of each word and the semantic valencies assigned to the words in the syntactic structure. Semantics of a single clause is represented by a *semantic predication*—a two-level tree with predicate ***p*** at the top and a number of actants at the bottom—where each node (predicate or actant) is a set of semantic markers. Such a semantic tree can be represented as a table—*frame*—taking into account that ***p*** is the head of the whole clause. The semantic frame of the sentence *Lingvisty otmetili psihologov na konferencii* ('Linguists have noticed psychologists at the conference') is shown in Table 1, while *U menya zvenit v ushah* ('It rings in my ears')—in Table 3.

Table 3. Semantic representation (frame) of the sentence *U menya zvenit v ushah* ('It rings in my ears')

Valency	***p***	***pos***	***loc***
Markers	to-sound	object somebody other-person principal (speaker)	object container container-small part part-of-human

For a compound (multi-clause) sentence several frames are constructed: one frame for each predicate—with co-reference or attributive links between the valencies of the frames. Further, the semantics of each tree is compared to the list of scripts (their input patterns) to determine the best tree and to suggest an emotional reaction for the F2 robot.

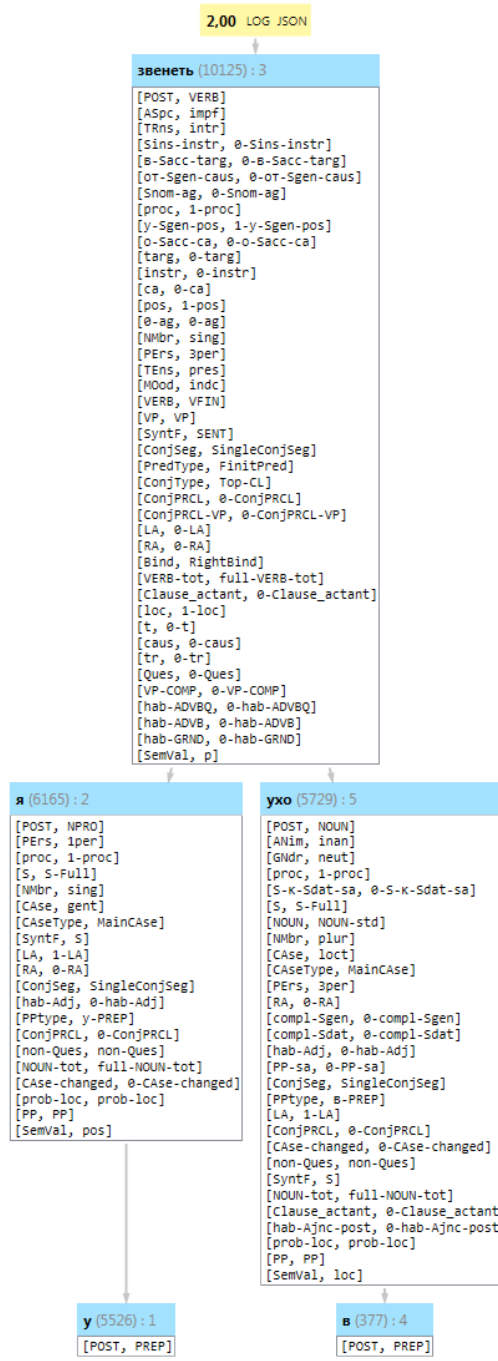


Fig. 2. Syntactic structure of the sentence *U menya zvenit v ushah* ('It rings in my ears'—*у меня звенит в ушах*)

2.3. Evaluation with scripts

To classify and evaluate the sentences, as well as to simulate the emotional dynamics for the F-2 robot, the parser relies on the set of *scripts*. Scripts are defined as *if-then* productions, a script is activated once an incoming semantic predication matches its' *if*-condition—*input pattern*. We consider an input pattern as an implementation of *frame*—a structure, similar to the semantic representation of a clause. For each incoming semantic predication, the parser calculates the distances with the input patterns of all the existing scripts. A tree with the highest similarity to the scripts is selected. We rely on the list of emotional scripts, represented in [Kotov, in press]. The list includes 13 scripts for negative situations: DANGER, APPROPR ("Appropriation"), SUBJV ("Subjectivity"—e. g. 'all he thinks about is himself!') etc., and 21 scripts for positive situations: CONTROL, CARE, COMFORT, ATTENTION (e. g. 'they all adore you!'), APPROVAL (e. g. 'you did it like a real man!') etc. By comparison of each incoming meaning (frame) to the script's input pattern the parser tries to classify a situation as a negative event—'this is dangerous', 'he is inadequate', or as a positive event—'they all adore me!'

Table 4. Distances with scripts and the possible responses for the utterance *Lingvisty otmetili psihologov na konferencii* ('Linguists have noticed psychologists at the conference')

Similarity	Script	Possible response
1.14	ATTENTION—gratitude	*Thank you for your support!*
1.14	ATTENTION	*It seems, everyone notices me!*
1.09	PLAN	*Did they mean that? They plan something evil!*
1.09	SUBJV	*They only think about this!*
1.01	UNUSUAL	*It's a whole new and unusual world!*
1.01	COMFORT	*It's nice to be here!*

As shown in Table 4, ATTENTION script is responsible for the "experience" of emotion, while "ATTENTION—gratitude" is responsible for a communicative reaction—the expression of gratitude, where the internal "expression" of emotion is not definitively required. In both scripts the robot associates itself with the ***pat*** valency ('a psychologist')—the most emotionally relevant position for the scripts. If any marker suggests a better reference to the robot (*Robot has noticed psychologists, Linguists have noticed the robot*)—the activation of scripts changes accordingly. Both attention scripts are sensitive to 'pay-attention' marker in ***p*** valency and 'somebody' markers in ***ag*** and ***pat*** valencies. Further scripts PLAN and SUBJV are responsible for negative reactions—they can be used to simulate a depressive mood or irony [Kotov, 2005]. Scripts UNUSUAL and COMFORT are sensitive to the situations where an agent acts inside or is entering some "magic world" or a "comfort place". They are sensitive to the actions, taking part in 'containers': UNUSUAL treats 'the conference' as a 'big and unusual container', like *walking in a wild forest*, while COMFORT considers the conference as a tiny and cozy container. As each semantic representation (frame) is associated with a script input pattern, each pair <*script input pattern, semantics*> serves as an illustration of *frame* and its *actual fillers*, like 'an idiot' = *John* or 'a cozy place' = *conference*.

Adverbs and adjectives may add additional markers to the predicate structures, changing their evaluation. For an utterance *Linguists always notice psychologists at the conferences*—SUBJV ("Subjectivity") becomes the leading script with the similarity 1.53—robot judges the 'linguists' as "limited" persons, who 'can do nothing else but noticing the psychologists'. This is an example of *emotionally sensitive marker*, as described in 2.1(b). Although the most active script is considered as the most relevant classifier of the situation and the best reaction for the robot, other scripts may be preferred for specific reactions: *They plan something evil!* reaction (PLAN) may be invoked to simulate robot's depressive state or irony [Kotov, 2009]. This ambiguity in the perception of text semantics may be used to study the architecture of consciousness—where one script (and the corresponding representation) constitutes the "accurate" understanding, while other scripts may form "ironical" or "hypothetical" understanding of the text [Kotov, 2017]. Different scripts may even suggest the selection of different syntactic trees, this option is not actually used, although it can be useful for computational simulation of humor, where the second semantics (tree) is used for the humorous semantic shift.

3. Facts database

Parser operates in a daily test mode: it collects through RSS and processes about 6,000 sentences per day from 25 information sources—15 media sites and 10 most popular LiveJournal blogs. The texts are analyzed by syntactic and semantic modules at a speed of 100 sentences per minute, the extracted semantic representations are stored in a database (PostgreSQL). As the parser is designed for the extraction of a specific set of scenarios (input patterns), the results of analysis cannot be directly compared to the results of other parsers: the analysis is not sensitive to the changes in the syntactic trees, not relevant to the input patterns. As we suggest, one of the possible extensions of the parser is the conversion of the extracted semantic representations to the input patterns of scenarios: up to now the parser distinguishes only the emotionally relevant representations. However, the extracted facts (as in Tables 1 and 3) can be converted to input patterns to recognize 'possible' or 'regular' situations. These situations do not offer specific communicative reactions to the robot (and thus their development is a side task for the project) but may extend the parser accuracy in a large number of the recognized situations.

The extracted shallow semantic representations can be easily stored and processed in a database. They offer diverse types of requests for information retrieval and classification.

Lexical analysis. For a given lexeme (head of a noun phrase or head of a verb phrase) it is possible to retrieve its participation in different valencies with other surrounding actants. As shown in Table 5, the analysis of a word distribution across valencies reveals the semantic features and polysemy of the word—where a single notion with certain polysemy may be treated as an active agent, patient, cause, location, time, source/target, or a trajectory. This approach may be used to extend the semantics of a particular word: e. g. add markers for a possible actor, time, 2d or 3d-location, etc.

Table 5. Participation of a lexeme ***snow*** in different valencies

Valency	Examples
ag (agent)	***Snow*** *falls.* ***Snow*** *melts under black feet.*
pat (patient)	*John has been raking* ***snow*** *with bare hands. Their children cannot eat* ***snow****.*
cont (content)	*He simply didn't notice the* ***snow****. It resembled a fire-breathing* ***snow****.*
instr (instrument)	*Traces would be drifted with the* ***snow****. His hands were frozen but he immediately rubbed them with* ***snow****.*
caus (cause)	*Everything is white because of the* ***snow****. The wreckage could happen due to the* ***snow****.*
foc (focus)	*We were lucky with the* ***snow****. Their ingression with the* ***snow*** *to the water was the reason of the crimson color.*
it (interpretation)	*He was surprised: it was a sweet* ***snow****! The construction has melted and turned into a simple* ***snow****.*
loc (location)	*Do they sleep right on the* ***snow****? People saw crows on the* ***snow****.*
t (time)	*After the* ***snow*** *in his dreams he knew, everything would be fine in the morning.*
src (source)	*An arrogant exclamation came out from the* ***snow****.*
targ (target)	*They threw rifles to the* ***snow****. He fell from the stairs and dug his head into the* ***snow****.*
tr (trajectory)	*He managed to get through the spring* ***snow****. He dragged the case in the* ***snow****.*

Analysis of a semantic/syntactic valency. For a given verb (head of a verb phrase) and its valency it is possible to study typical fillers—like *what do people drink?* (patient for a verb *drink*—see Table 6) or *who is usually a patient of violence?* (patient of any violence verb). This aggregation may serve as a basis for question answering—for questions to a specific participant of a situation, as well as to extend lexical semantics: e. g. add markers of a possible 'drink' to the fillers of the frame.

Table 6. Fillers of the ***pat*** valency for diverse 'drink' predicates

pat of 'drink' predicates	Number of cases
tea	144
water	78
wine	72
coffee	64
beer	51
what	49
vodka	40
blood	31
something	19
champagne	17

[2] Wrong analysis of a clause *drink* [*is a*] *sin*.

pat of 'drink' predicates	Number of cases
milk	14
glass	13
bottle	11
drink	10
cognac	10
it	10
sin[2]	10
cocktail	9
liquid	9

Search and aggregation of semantic frames. The annotation of the valencies with semantic markers makes it possible to search for all the facts (semantic predications) corresponding to a given pattern. For example, in the cases of extremism, where 'a person causes harm to someone', named after his/her 'nationality' or 'profession'—as in the following examples:

(1) *His girlfriend has beaten the driver with an unidentified object.*

(2) *Following the investigation records, the businessman has shot the director of the plant.*

(3) *Husband has also started to beat the policemen.*

The aggregation of specific examples, meeting the requested pattern may be further used to extend the accuracy of the input patterns. In this sense, the parser represented here is *frame*-centric: frames (input patterns of scripts) are used to select a tree and each word meaning, further, frames can be used to extend lexical semantics, new frames can be designed basing on the clustering of the extracted meanings, and the existing frames may be refined by the new examples falling within the same script.

References

1. *Blakar, R. M.* (1979), Language as a means of social power. In: Pragmalin-guistics, J. Mey (ed.). The Hague-Paris, Mouton, pp. 131–169.
2. *Clore, G. L., & Ortony, A.* (2000), Cognition in Emotion: Always, Sometimes, or Never? In R. D. Lane & L. Nadel (Eds.), Cognitive Neuroscience of Emotion: Oxford Univ. Press, pp. 24–61.
3. *Fillmore, C. J.* (1968), The Case for Case. In Universals in linguistic theory. New York: Holt, Rinehart & Winston, pp. 1–68.
4. *Gladkiy, A. V.* (1985), Syntactic Structures of Natural Language in Computer-Aided Communication Systems [Sintaksicheskie struktury estestvennogo yazyka v avtomatizirovannyh sistemah obshcheniya], Nauka, Moscow.
5. *Golovinskaya, M. Ya.* (2004), Hidden Hyperbole as a Manifestation and Justification of Verbal Aggression. The Hidden Meanings: Word. Text. Culture [Skrytaya giperbola kak proyavlenie i opravdanie rechevoj agressii Sokrovennye smysly: Slovo. Tekst. Kul'tura], Yazyki slavyanskoj kul'tury, Moscow, pp. 69–76.

6. *Granovsky, D. V., Bocharov, V. V., Bichineva, S. V.* (2010), The OpenCorpora: Principles and Prospects [Otkrytyj korpus: principy raboty i perspektivy], Computational Linguistics and Development of Semantic Search on the Internet: Proceedings of the Scientific Workshop of the XIII All-Russian Joint Conference "Internet and modern society" [Komp'yuternaya lingvistika i razvitie semanticheskogo poiska v Internete: Trudy nauchnogo seminara XIII Vserossijskoj ob"edinennoj konferencii «Internet i sovremennoe obshchestvo»], St. Petersburg, p. 94.
7. *Kotov, A.* (2017), A computational model of consciousness for artificial emotional agents. Psychology in Russia: State of the Art, 10 (3), pp. 57–73.
8. *Kotov, A.* (2005), Application of Psychological Characteristics to D-Script Model for Emotional Speech Processing J. Tao, T. Tan, and R. W. Picard (Eds.) ACII 2005, LNCS 3784. Berlin, Heidelberg: Springer-Verlag, pp. 294–302.
9. *Kotov, A.* (2009), Accounting for irony and emotional oscillation in computer architectures Proceedings of International Conference on Affective Computing and Intelligent Interaction ACII 2009. Amsterdam: IEEE, pp. 506–511.
10. *Kotov, A.* (in print), Mechanisms of speech influence. M.: Kurchatov Institute.
11. *Kotov, A., Zinina, A., & Filatov, A.* (2015), Semantic Parser for Sentiment Analysis and the Emotional Computer Agents Proceedings of the AINL-ISMW FRUCT 2015 (pp. 167–170).
12. *Lyashevskaya O.* (2010), Bank of Russian constructions and valencies // LREC 2010. Malta, Valletta, May 19–21.
13. *Melčuk, I. A.* (1999), The Theory of Linguistic Models "Meaning ⇔ Text" [Opyt teorii lingvisticheskih modelej «SMYSL ⇔ TEKST»], Jazyki russkoj kultury, Moscow.
14. *Minsky, M. L.* (1988), The Society of Mind. New-York, London: Touchstone Book.
15. *Shvedova, N. Yu.* (1998) The Russian Semantic Dictionary [Russkij semanticheskij slovar'. Tolkovyj slovar', sistematizirovannyj po klassam slov i znachenij], Azbukovnik, Moscow.
16. *Yeh, W., & Barsalou, L. W.* (2006), The situated nature of concepts. American Journal of Psychology, 119(3), pp. 349–384.

БАЗА ДИСКУРСИВНЫХ ПРИЗНАКОВ СЛОВОРАЗДЕЛА В УСТНОЙ РУССКОЙ РЕЧИ: СТРУКТУРА, СОСТАВ И ОПЫТ ПРИМЕНЕНИЯ

Кривнова О. Ф. (okrivnova@mail.ru),
Смирнова О. С. (kisaolga@mail.ru)

Московский государственный университет
имени М. В. Ломоносова, Москва, Россия

Ключевые слова: устная речь, просодическое членение, словораздел, сегментирующая сила, просодический шов, иерархия, речевой корпус, база данных, синтаксический, статистический, инструментальный анализ

A DATABASE OF WORDBREAKS DISCURSIVE FEATURES IN RUSSIAN ORAL SPEECH: THE STRUCTURE, COMPOSITION AND APPLICATION

Krivnova O. F. (okrivnova@mail.ru),
Smirnova O. S. (kisaolga@mail.ru)

Moscow State University, Moscow, Russia

The paper discusses the most important results of the project “Hierarchy of prosodic phrasing in spoken language: controlling factors and means of realization”. The project was aimed at expanding the empirical base of phrasal prosody researches, which inadequacy is marked in many scientific areas: discourse theory, syntax, intonational phonology, general phonetics, speech synthesis and recognition etc. The introduction provides a brief description of the study background and formulates the tasks which were necessary to solve for the ultimate goal of the project planned for 3 years of implementation. The first section describes the characteristics of speech corpora created in the the project for construction of a complex, linguistic-prosodic database required for the study and modeling of prosodic phrasing in Russian speech, which takes into account, if possible, all controlling factors and means of realization. The second section is devoted to the description of the structure and composition of wordbreaks’ discursive features database (BDF), obtained on the basis of annotated, prosodically graduated and acoustically analyzed speech corpora. It should

be noted the universality and flexibility of the format and structure of the database as a computer resource, freely admitting to extend its feature set and to detail their parametric characteristics. The third section illustrates as the BDF application for theoretical and statistical modelling of inter-level correlations “syntax—linguistic prosody” in both directions and “linguistic prosody and speech signal (acoustic speech)” in both directions. The conclusion summarizes the results of research and discusses some promising directions for further studies on relevant topics.

Key words: phonetics, spoken language, prosodic phrasing, wordbreak’ strength, prosodic break depth, hierarchy, database, syntactic, statistical and instrumental analysis

1. Введение

Цель проекта, о котором идет речь в настоящем докладе, заключалась в исследовании иерархии просодического членения (ПЧ) русской звучащей речи с целью получения новых и статистически надежных данных о природе этого явления, фонетической реализации и контролирующих факторах: коммуникативно-синтаксических, фонетических и физиологических. Важным моментом теоретической установки исследования является признание слова основной рабочей единицей как при порождении, так и при восприятии любого текста, и письменного, и устного. С точки зрения возможностей ПЧ любая граница между словами (словораздел) в тексте имеет определенный сегментирующий *потенциал*, который может реализоваться с разной вероятностью и с разной силой в зависимости от типа дискурса, контекстных условий и различных контролирующих факторов локальной и глобальной природы. Носитель языка, даже не имеющий специального лингвистического образования, согласится с тем, что соседние слова в тексте в разной степени связаны между собой по смыслу, синтаксически и даже фонетически. Этот факт можно интерпретировать как признание разной сегментирующей силы (глубины) словоразделов. В *письменном* тексте в качестве формальных показателей сегментирующей силы словоразделов (wordbreak strength) выступают знаки препинания: их наличие/отсутствие и тип. Знаки препинания не только членят текст на когерентные фрагменты, но и указывают в определенной степени на их иерархический статус. В *устной* речи аналогичную функцию выполняют просодические средства: паузы, переломы тона и другие фонетические явления на граничных участках соседних слов в последовательности. Членение звучащего текста с помощью просодических средств осуществляется говорящим в соответствии с общими принципами фонетической организации речи и с учетом смысловой и синтаксической структуры текста. Просодически маркированные словоразделы между фразовыми просодическими составляющими образуют просодические швы (разрывы) в звучащем тексте, что хорошо отражает англоязычный термин «prosodic break». Естественно предположить, что внутренняя иерархия ПЧ на фразовом уровне находит отражение в разной глубине («силе») просодических швов (далее ПШ), которая создается использованием разных просодических средств между

и на краях просодических составляющих. В интонационной фонологии многие исследователи разделяют точку зрения, согласно которой иерархический статус просодической составляющей однозначно соответствует глубине ПШ, завершающего эту составляющую. Это положение т. н. строгой поуровневой гипотезы (Strict Layered Hypothesis SLH) разделяется, однако, не всеми интонологами, но, к сожалению, никогда не проверялось экспериментально на сколько-нибудь представительном речевом материале, см. об этом [Ladd 1986; Ladd, Campbell 1991; Sanderman 1996; Selkirk 1984]. В то же время нельзя не отметить, что количественная оценка сегментирующей силы словоразделов, или, что то же самое, глубины ПШ в озвученном речевом высказывании, создает основу для реконструкции иерархической структуры его просодических составляющих.

В отечественной лингвистике впервые обратил внимание на ПЧ и его особую природу академик Л. В. Щерба. Он писал: «В европейских языках (а вероятно и во многих других) самым могучим средством выражения связи между словами и группами слов является „интонация“, „фразировка“ в самом широком смысле слова» [Щерба 1915]. Щерба обозначил практически все отличительные особенности этого явления, которые в настоящее время являются объектом исследования во многих работах по фразовой просодии, однако не описаны и не объяснены полностью ни для одного из европейских языков. Это относится, в частности, и к иерархической природе ПЧ. Щерба об этом писал так: «синтагмы (минимальные единицы ИЧ) могут объединяться в группы высшего порядка с разными интонациями и в конце концов образуют фразу — законченное целое, которое может состоять из группы синтагм, но может состоять и из одной синтагмы, и которое нормально характеризуется конечным понижением тона» [Щерба 1955]. В этой же работе приведены авторские транскрипции русского стиха, где используются 4 маркера для фразовых ПШ разной глубины. Приведенные примеры дают основания считать, что автором использовалась пятибалльная количественная шкала: 0, 1, 2, 3, 4. Много текстовых примеров с интроспективной разметкой ПШ содержится и в книге Р. И. Аванесова «Русское литературное произношение» [1972]. В приводимых им примерах используются пять маркеров глубины ПШ, т. е. фактически шестибалльная количественная шкала.

Результаты перцептивных экспериментов по оценке сегментирующей силы словоразделов на материале разных языков (английского, нидерландского и русского) свидетельствуют о поведенческой устойчивости оценок, полученных с использованием количественной шкалы, различающей не более 5 уровней глубины ПЧ. Это позволяет предположить, что данная шкала отражает какие-то универсальные свойства перцептивных ощущений человека в анализируемом фонетическом пространстве. Для многих практических задач достаточно учитывать этот результат и даже сознательно использовать более крупную трехуровневую шкалу, как это советует делать А. Сандерман [Sanderman 1996] на начальной стадии технологических разработок в области синтеза и распознавания речи. Однако для подтверждения указанной гипотезы с общефонетических научных позиций необходимо увеличить как количество и разнообразие рассматриваемых языков, так и число испытуемых, привлекаемых к ее верификации. Подробнее см. об этом [Кривнова 2015].

Исходя из изложенных теоретических установок, в нашем проекте были решены следующие конкретные задачи:

1. Создан просодически размеченный речевой корпус устной русской речи на базе известной разметочной системы TOBI (Tone and Break Indices) с привлечением фонетистов-экспертов и обычных носителей языка в качестве дикторов и аудиторов.
2. На основе комплексного лингво-акустического анализа текстового и звукового материала корпуса построена база данных текстовых характеристик каждого словораздела (далее БДС) в корпусе (вектор признаков) с целью дальнейшего анализа зависимости между этими признаками и сегментирующей силой словораздела с включенным в него ПШ разного типа и глубины. БДС реализована в двух форматах — как электронная таблица EXCEL и как таблица статистического пакета STATISTICA.
3. В пакете STATISTICA производилась статистическая обработка данных на всех этапах проекта и анализ требуемых корреляций. Примеры и результаты решения некоторых задач кратко обсуждаются в последующих разделах настоящей работы.

2. Материал и методика исследования

Учитывая многообразие форм устного дискурса, многослойность и многоаспектность проблемы ПЧ, речевой материал проекта был ограничен русскими прозаическими текстами, озвученными в режиме чтения, с некоторыми дополнениями в виде текстов, специально озвученных для целевых фонетических экспериментов. Основной массив корпуса включает два фрагмента из русских прозаических текстов:

1. фрагмент из повести И. Грековой «Кафедра» (2700 словоупотреблений = словоразделов в чтении 2-х непрофессиональных дикторов, далее текст ИГ)
2. коллекция прозаических текстов из книги Р. И. Аванесова «Русское литературное произношение» (2500 словоупотреблений = словоразделов в чтении 3-х непрофессиональных дикторов, далее текст РИА). Дополнительный массив корпуса включал небольшой современный рассказ в прочтении 10 дикторов (5 мужчин и 5 женщин). Этот массив использовался в целевом исследовании речевого дыхания как фактора ПЧ [Кривнова 2017]. Во всех случаях запись производилась на качественной цифровой аппаратуре в студийных условиях.

Тексты основного массива были выбраны для корпуса не случайно: текст ИГ входит в набор текстов, имеющих синтаксическую разметку в формате анализатора ЭТАП-3, а тексты РИА содержат авторскую разметку просодического членения. В то же время при включении этих текстов в состав корпуса потребовалась унификация сопутствующих им лингво-просодических данных.

Фонетическая составляющая основного массива корпуса и БДС включает перцептивную разметку, в которой каждому словоразделу поставлен

в соответствие количественный субъективный показатель его сегментирующей силы, соответственно глубины ПШ. Разметка осуществлялась с использованием фиксированной 5-ти балльной шкалы, согласованной с результатами перцептивных экспериментов, проведенных в рамках проекта с привлечением фонетистов-экспертов и обычных носителей языка в качестве аудиторов. Более подробно см. об этом [Смирнова 2017] и [Кривнова и др. 2018]. Кроме этого, была также произведена и верифицирована полная паузальная разметка как основного, так и дополнительного массивов корпуса с одновременным измерением следующих временных показателей: длительности паузы на словоразделе, длительности предпаузального и постпаузального речевых фрагментов.

Помимо описанной просодической разметки, в состав фонетических признаков БДС входят также акустические характеристики (длительность, интенсивность, ЧОТ, спектральные показатели коартикуляции) звуковых сегментов в окрестности каждого словораздела. Для получения этих характеристик была произведена ручная сегментация пограничных областей для каждого словораздела и автоматический акустический анализ выделенных звуковых сегментов.

Синтаксическая составляющая корпуса и БДС формировалась постепенно, и вопрос о выборе синтаксического формализма, наиболее адекватного для наших задач, до сих пор требует доработки. Текст ИГ, исходно размеченный помощью анализатора ЭТАП-3 в формате деревьев зависимостей для каждого предложения текста, был программно переведен в скобочную форму синтаксического представления предложения в терминах непосредственных составляющих. Скобочная форма более удобна для представления синтаксической информации в табличной форме и позволяет вычислять и использовать простые и осмысленные для синтаксической интерпретации ПЧ признаки, как то: плотность (сгущение) скобок на словоразделе, соотношение закрывающих и открывающих скобок, тип объемлющей составляющей по обе стороны от словораздела и т. п. Кроме того, для ИГ были произведены две альтернативные ручные разметки синтаксиса. В одной из них использовался метод [Гаспарова-Скулачевой 2004], который применяется для исследования глубины синтаксических границ в структуре стихотворного текста. Этот тип разметки основан на классической синтаксической теории членов предложения и эмпирически установленной иерархии силы синтаксической связи между словами, разделенными словоразделом. Второй тип ручной разметки можно условно назвать синтактико-пунктуационным: для получения данных о влиянии типа синтаксической составляющей на ПЧ было выделено 12 типов потенциально релевантных синтаксических структур, а затем была проведена ручная разметка текста маркерами таких составляющих. Во многих случаях границы таких составляющих в тексте маркируются в соответствующих словоразделах знаками препинания, согласно действующим в русском языке пунктуационным правилам[1],

[1] С точки зрения реальных механизмов, контролирующих появление ПШ в тексте, естественно предполагать, что диктор при озвучивании текста ориентируется в известной степени на знаки препинания, а также и на стоящий за ними синтаксис.

однако далеко не всегда. Таким образом, синтаксический фактор влияет на паузы не только опосредованно через знаки препинания, но должен учитываться и независимо от них. В приводимой ниже таблице можно видеть ПШ глубины 1 на словоразделе без ЗП, но с некоторым сгущением синтаксических скобок. При этом один из таких ПШ реализован без паузы, что достаточно характерно для ПШ малой глубины.

Тексты РИА были размечены вручную с выделением наиболее значимых и крупных синтаксических границ: границы текста, конец абзаца внутри текста, конец самостоятельного предложения внутри абзаца, конец элементарной клаузы внутри самостоятельного предложения, границы пояснительных и сравнительных оборотов внутри элементарной клаузы. Значимость этой разметки для ПЧ также исследовалась статистически.

3. Структура и состав комплексной лингво-просодической БДС в звучащем тексте

В структуре разработанного макета БДС имеются четыре логические зоны: текстовая зона, зона левого контекста словораздела, центральная зона собственно словораздела и зона его правого контекста. Каждая зона имеет в своем составе определенный набор характеристик (признаков разного типа). Текстовая зона представляет собой пословную орфозапись целевого текста, где каждое графическое слово занимает отдельную строку[2] таблицы (БДС), номер которой соответствует порядковому номеру словораздела, следующего за словом с тем же порядковым номером. Зона левого контекста в настоящее время включает только частеречный признак слова[3], стоящего непосредственно перед словоразделом; планируется ввести информацию о типе объемлющей синтаксической составляющей, граничащей слева с данным словоразделом. Кроме того, в этой же зоне фиксируется несколько фонетических признаков: длительность/длина речевого фрагмента перед словоразделом, с разными возможностями выбора начальной границы фрагмента. Здесь же приводятся акустические показатели звуковых сегментов конечного участка просодической составляющей перед словоразделом. Центральная зона собственно словораздела включает информацию о знаке препинания на словоразделе, формальные синтаксические «скобочные» признаки — общее количество скобок НС, соотношение закрывающих и открывающих скобок. Из фонетических признаков — длительность паузы и наличие в ней включенного вдоха. Состав признаков зоны правого контекста словораздела, в целом, зеркально аналогичен составу зоны левого контекста.

Заметим, что реальная физическая структура БДС по ряду технических причин несколько отличается от описанной выше логической. Это связано

---

[2] Для упрощения БД некоторые последовательности из «мелких» служебных слов с союзным, модальным, наречным значением условно считались одним графическим словом, т.е. устойчивым выражением, эквивалентным слову (УВР), (*будто бы*, *потому что*, *в течение*, *или же* и др.) [Рогожникова 2003].

[3] Базовый список различаемых частей речи был взят из НКРЯ 2003–2005.

с тем, что поля текущего варианта БДС, описывающие различные признаки/атрибуты словораздела, добавлялись в базу в рабочем порядке, по мере выполнения соответствующих исследований (перцептивная просодическая разметка, преобразование древесной синтаксической структуры в линейную скобочную запись НС, другие виды синтаксической, частеречной и фонетической разметки). В таком виде она достаточно удобна как для визуального восприятия, так и для статистической обработки.

Приведенная ниже в качестве иллюстрации таблица сокращена: в ней показаны только те атрибуты текущей версии БДС, которые по результатам статистической обработки данных оказались наиболее значимыми для моделирования ПЧ. Например, из 12-ти факторов синтактико-пунктуационной разметки показаны только пять, а из разметки по методу Гаспарова-Скулачевой только основной тип связи, имеющий максимальную предикторную силу для выбора типа и локализации ПШ.

Выбор электронных форматов для разрабатываемой БДС (в виде рабочей книги EXCEL и рабочей таблицы статистического пакета) не случаен. Практически все пользователи умеют в какой-то степени работать с теми или иными версиями электронных таблиц и располагают их лицензионными копиями. Кроме того, информация из EXCEL импортируется во все универсальные статистические пакеты, так что использование именно пакета STATISTICA не принципиально и объясняется только привычным удобством его интерфейса и программирования в среде пакета. Вариант БДС в формате статпакета является более мобильным, рабочим, предназначенным для вычислений, программного анализа данных. Вариант базы в формате EXCEL предназначен скорее для исследователей-нематематиков, он допускает более красивое оформление и более удобен для пилотного визуального анализа, который предшествует построению математических моделей ПЧ с использованием вектора признаков словораздела, имеющих различную природу. Данные в этом варианте текущей версии БДС защищены, он является эталонным. Оба формата могут использоваться также для получения «линейных» орфотекстов с различными типами разметок (паузальной, с ПШ, синтаксической, частеречной и пр.).

4. Применение лингво-просодической БДС для исследования ПЧ в звучащем тексте

Ниже в иллюстративной таблице БДС выделены две строки (7 и 13), показывающие возможности визуализации и качественного анализа данных, присутствующих в базе, а именно: появление в звучащем тексте ПШ глубины 2 и 4 и, соответственно, относительно короткой и длинной пауз. На этих примерах хорошо видно влияние на вероятность появления ПШ таких факторов как знак препинания, «сгущение» синтаксических скобок, часть речи (существительное) в левом контексте словораздела, элементы синтаксической разметки (маркеры границы финитной клаузы, бессоюзного сочинения, левой границы оборота), слабый тип (cob, f) синтаксической связи. В последней строке таблицы видно также, как метка конца абзаца усиливает эффект воздействия прочих

факторов, давая глубину ПШ, равную 5. Эти неформальные наблюдения подтверждаются статистическим анализом зависимостей между факторами.

Помимо признаков основных разметок, база содержит ряд вычисленных характеристик словораздела, влияние которых на выбор локализации и глубины ПШ может быть значительным, а использование более удобным и убедительным, чем использование «сырых» данных. К ним относятся, например, укрупненные шкалы знаков препинания, производные признаки скобочной разметки (относительные сгущения, смена направления скобок, различные характеристики сложности синтаксической структуры предложения и др.).

На всех этапах проекта проводились статистическая обработка полученного материала, начиная с просодической разметки текста (столбец БДС «ПШ»), которая является результатом нетривиальной статистической обработки данных перцептивных экспериментов [Смирнова 2017]. Следует отметить устойчивость результата, значительное совпадение разметок начального фрагмента текста ИГ, полученных на разных этапах проекта при разном составе групп аудиторов и несколько различном дизайне эксперимента (ужесточении правил прослушивания и записи результата).

Почти все проанализированные нами показатели являются номинальными, иногда допускающими также ранговую интерпретацию (ПШ, ранжированные ЗП, тип синтаксической связи в иерархической «стиховедческой» разметке) и использование соответствующих методов непараметрической статистики. Основным статистическим аппаратом на начальном этапе исследований являлся анализ таблиц сопряженных признаков — проверка гипотез независимости признаков и вычисление различных мер связи, в том числе асимметричных, «прогнозных», на основании чего были выделены наиболее значимые для ПЧ факторы-атрибуты в разных зонах БДС.

К настоящему времени единственным проанализированным числовым, непрерывным, атрибутом базы является длительность паузы. Вид выборочной функции распределения позволяет предположить, что распределение длительности пауз является смесью нескольких распределений со сравнительно небольшой вариацией и выраженным наиболее вероятным значением. Даже очень грубое формальное разделение выборки длительностей пауз «по глубине ПШ» позволило не только продемонстрировать ожидаемую зависимость этих показателей, но и построить диапазоны длительностей, соответствующих глубине ПШ, в которые паузы попадают с вероятностью более 80%, получить оценки наиболее вероятных значений пауз внутри этих диапазонов. Эти результаты хорошо согласуются с результатами О. Ф. Кривновой, полученными ранее другими методами [Кривнова 2015; Смирнова 2017].

Нами предпринималась и попытка формальной категоризации длительности пауз. Переход от числовой шкалы к укрупненной номинальной дает классы с достаточно устойчивыми центроидами, близкими к наиболее вероятным значениям внутри классов («нет паузы». «минимальная» — в среднем 100–200 мс, «короткая» — 400–500, «средняя» — 700–800, «большая, длинная» — 1000–1200, «максимальная» — порядка 2000 мс; паузы из максимального класса часто имеют особую природу: связаны с переворачиванием

страницы или с оговорками диктора. Приведенные значения категориальных длительностей сильно округлены и, безусловно, зависят от специфики текста и темпа речи диктора. Исследование этой зависимости может быть предметом отдельного исследования, но здесь стоит отметить, что анализ рангового распределения для полученной нами классификации показывает неплохое согласие с законом Ципфа-Мандельброта[4], т. е. классификация может рассматриваться как «естественная».

Ранжированная временная шкала может оказаться полезной при построении моделей ПЧ, при ее использовании максимально очевидна зависимость между категорированной длиной паузы и глубиной ПШ, а также некоторыми другими факторами.

5. Заключение

В докладе описан экспериментальный макет базы дискурсивных признаков словораздела, разработанный для исследования просодического членения звучащей русской речи. Чтобы стать эффективным инструментом моделирования ПЧ, созданная база должна быть расширена в нескольких направлениях. Так, оставаясь в рамках того же текстового материала, в фонетическую зону БДС нужно включить информацию о наличии, типе и степени выделенности тонального акцента на словах, разделенных словоразделом, а в зону собственно словораздела — информацию о наличии в словоразделе особых фонетических явлений: вдохов разного типа и глубины, признаков ларингализации и, возможно, других особенностей фонации. При наличии нескольких прочтений одного и того же текста одним и тем же или разными дикторами, с разными установками на степень выразительности чтения, нужно предусмотреть возможность оперативного и удобного включения в БДС характеристик этих прочтений и их статистическую обработку. Это позволит уточнить понятие и содержание разграничения между обязательным и факультативным ПЧ, которое достаточно часто декларируется в интонационной литературе. Наконец, оставаясь в рамках предложенной структуры БДС, нужно предусмотреть и опробовать возможности введения и анализа информации о ПЧ в других типах монологического дискурса: поэтической, публичной, научной театральной и других видов спонтанной речи с разной степенью подготовленности и т. д. В методическом плане необходимы дальнейшие исследования зависимости ПЧ от вектора признаков словораздела методами многомерной классификации/регрессии с целью построения формальных моделей адекватного прогноза ПЧ: локализации и глубины ПШ и их фонетической реализации. Такие модели необходимы не только для расширения фонетических знаний в области фразовой просодии, но и для создания компьютерных систем качественного синтеза и распознавания русской речи в разных дискурсивных контекстах.

---

[4] Закон Ципфа-Мандельброта утверждает линейную зависимость логарифмов частоты и ранга элементов при их ранжировании по убыванию частоты.

Литература

1. *Аванесов Р. И.* (1972) Русское литературное произношение. Просвещение, М.
2. *Гаспаров М. Л., Скулачева Т. В.* (2004) Статьи о лингвистике стиха. М.: Языки славянской культуры.
3. *Кривнова О. Ф.* (2015). Глубина просодических швов в звучащем тексте (экспериментальные данные) // Компьютерная лингвистика и интеллектуальные технологии. Материалы ежегодной международной конференции «Диалог». М., РГГУ, 2015, вып. 14, т. 1., сс. 326–338.
4. *Кривнова О. Ф.* (2017)Фонетические характеристики дыхательных пауз с разной текстовой локализацией // Компьютерная лингвистика и интеллектуальные технологии. Материалы ежегодной международной конференции «Диалог». М., РГГУ, 2017, вып. 14, т.1., сс. 315–325.
5. *Кривнова О. Ф., Князев С. В., Смирнова О. С.* (2018) Интонационное членение и сегментируюшая сила словоразделов в звучащем тексте (данные перцептивного эксперимента)// сб. Фонетика, М. ИРЯ РАН, в печати.
6. *НКРЯ* — Национальный корпус русского языка 2003–2005: Результаты и перспективы. М.
7. *Рогожникова Р. П.* (2003) Толковый словарь сочетаний эквивалентных слову. М.
8. *Смирнова О. С.* (2017) Статистический анализ результатов перцептивного оценивания глубины просодических швов в русском звучащем тексте // Компьютерная лингвистика и интеллектуальные технологии. Материалы ежегодной международной конференции «Диалог». М., РГГУ, 2017, электронная публикация.
9. *Shcherba L. V.* (1915) Восточнолужицкое наречие, Пгр.
10. *Щерба Л. В.* (1955) Фонетика французского языка. М.

References

1. *Avanesov R. I.* (1972) Russian Literary Pronunciation [Russkoe literaturnoe proiznoshenie], Education, M.
2. *Gasparov M. L., Skulacheva T. V.* (2004) Papers on verse linguistics [Stat'ji o lingvistike stiha] The languages of Slavic cultures, M.
3. *Krivnova O. F.* (2015) The depth of prosodic breaks in spoken text (experimental data) [Glubina prosodicheskih shvov v zvuchaschem tekste (eksperimental'nyje dannyje)] Computer linguistics and intellectual technologies. Proceedings of the annual international conference "Dialogue" [Komp'juternaja lingvistika I intellektual'nyje tehnnologii. Materialy jezhegodnoj mezhdunarodnoj konferentsii 'Dialog'] M., RGGU, v. 14, t. 1, pp. 326–338.
4. *Krivnova O. F.* (2017) Phonetic characteristics of breathing pauses with different text localization [Foneticheskije harakteristiki dyhatel'nyh payz s raznoj tekstovoj lokalizatsiej] Computer linguistics and intellectual technologies. Proceedings of the annual international conference "Dialogue" [Komp'juternaja lingvistika

I intellektual'nyje tehnologii. Materialy jezhegodnoj mezhdunarodnoj konferentsii 'Dialog'] M., RGGU, v. 16, t.2, pp. 207–220.
5. *Krivnova O. F., Knyazev S. V., Smirnova O. S.* (2018) Prosodic phrasing and word-break' strength in oral text (experimental data) [Intonatsionnoje chlenenije I segmentirujuschaja sila slovorazdelov v zvuchaschem tekste (eksperimental/nyje dannyje)] In coll. Phonetics [Fonetika] M. IRIA RAN.
6. *Ladd R.* (1986) Prosodic phrasing: a case of recursive prosodic structure. Phonology Yearbook3, pp. 311–340.
7. *Ladd B., Campbell D. R.* (1991) Theories of prosodic structure: evidence from syllable duration // Proc. of the 12th Congress of Phonetic Sciences, Aix-en-Provance, France, pp. 290–293.
8. *Sanderman A.* (1996) Prosodic Phrasing (production, perception, acceptability and comprehension). Eindhoven.
9. *Silverman K., Beckman M., Pitrelli J., Ostendorf M., Wightman C., Price P., Pierrehumbert J., Hirshberg J.* ToBI: A standart for labelling English prosody // Proc, of the 1992 International Conference on Spoken Language Processing (ICSLP), pp. 867–870.
10. *Selkirk E.* (1984) Phonology and syntax: the relation between sound and structure, MIT, Cambridge.
11. *Smirnova O. S.* (2017) Statistical analysis of the results of prosodic breaks' perceptual evaluation in spoken Russian text [Statisticheskij analiz rezul'tatov pertseptivnogo otsenivanija glubiny prosodicheskih shvov v russkom zvuchaschem tekste] // Computer linguistics and intellectual technologies. Proceedings of the annual international conference "Dialogue" [Komp'juternaja lingvistika i intellektual'nyje tehnikigii. Materialy jezhegodnoj mezhdunarodnoj konferentsii 'Dialog'] M., RGGU, 2017. Electronic publication.
12. *Shcherba L. V.* (1955) Phonetics of the French language [Phonetika frantsuzskogo jazyka] Publishing house of foreign literature, M.

Таблица 1. База дискурсивных признаков словораздела в устной русской речи (фрагмент)

Номер словораздела	Текст ИГ в послловной орфозаписи и одновременно левый контекст словораздела	ЗП				СИНТАКСИС — ЭТАП				ПШ	ПАУЗЫ			ЧАСТИ РЕЧИ		Альтернативная синтаксическая разметка						
		Знак препинания	ЗП+	ЗП 0/1	ЗП 0/1/2/3	Число закрывающих]]	Общее число откр [[на сл/разделе	Всего][скобок на словоразделе	Число [[следующего предложения	ПШ по согласованной группе аудиторов	Абзац	Длительность паузы (мс)	Длительность речевого фрагмента (мс)	Краевые сегменты	Часть речи левый контекст (ЛК)	Правый контекст (ПК)	Граница финитной клаузы	Бессоюзное сочинение	Союзное сочинение	Пояснительный оборот	Левая граница в обороте	Иерархический тип синтаксической связи
0						0	2	2	2							сущ						
1	Люда	–	–	0	0	0	1	1	0	0	0	0,0		–	сущ	сущ	0	0	0	0	0	a
2	Величко	–	–	0	0	2	0	2	0	1	0	**128,0**	816,0	–	сущ	глаг	0	0	0	0	0	p
3	родилась	–	–	0	0	0	1	1	0	0	0	0,0		–	глаг	союз	0	0	1	0	0	c
4	и	–	–	0	0	0	1	1	0	0	0	0,0		–	союз	глаг	0	0	0	0	0	a
5	росла	–	–	0	0	0	1	1	0	0	0	0,0		–	глаг	предл	0	0	0	0	0	m
6	в	–	–	0	0	0	1	1	0	0	0	0,0		–	предл	сущ	0	0	0	0	0	a
7	провинции	,	зпт	1	1	1	1	2	0	2	0	**540,0**	1615,0	–	сущ	предл	0	0	0	1	1	cob
8	в	–	–	0	0	0	0	1	0	0	0	0,0		–	предл	числ	0	0	0	0	0	a
9	одном	–	–	0	0	0	0	1	0	0	0	0,0		–	числ	предл	0	0	0	0	0	i
10	из	–	–	0	0	0	0	2	0	0	0	0,0		–	предл	прил	0	0	0	0	0	a
11	среднерусских	–	–	0	0	1	1	2	0	1	0	0,0		–	прил	прил	0	0	0	0	0	a
12	неперспективных	–	–	0	0	1	0	1	0	0	0	0,0		–	прил	сущ	0	0	0	0	0	a
13	городков	.	тчк	1	3	8	2	10	2	4	0	**1114,0**	2900,0	–	сущ	числ	1	0	0	0	0	f
14	Одна	–	–	0	0	1	0	1	0	0	0	0,0		–	числ	сущ	0	0	0	0	0	a
15	фабрика,	,	зпт	1	1	0	2	2	0	2	0	**353,0**	858,0	–	сущ	прил	1	1	0	0	0	c
16	лесопильный	–	–	0	0	1	0	1	0	0	0	0,0		–	прил	сущ	0	0	0	0	0	a
17	завод ……	,	зпт	1	1	0	2	2	0	2	0	**358,0**	1109,0	–	сущ	прил	1	1	0	0	0	c
115	пиявки.	.	тчк	1	3	2	4	2	0	5	1	**3174,0**	2316,0	–	сущ	прил	1	0	0	0	0	f

Наиболее значимые (по результатам предварительного анализа) для формирования глубины ПШ факторы выделены жирным шрифтом.

МЕНТАЛЬНЫЕ ПРЕДИКАТЫ 2-ГО ЛИЦА В МЕТАТЕКСТОВЫХ КОНСТРУКЦИЯХ[1]

Кустова Г. И. (galinak03@gmail.com)

Институт русского языка им. В. В. Виноградова РАН; Московский педагогический государственный университет; Москва, Россия

В работе рассматриваются метатекстовые (вводные) конструкции с ментальными глаголами во 2-м лице. Показано, что если пропозиции, ассоциированные с вводными словами 1-го лица (*думаю*; *боюсь*; *знаю* и т. д.) и 3-го лица (*считают* и т. п.) принадлежат говорящему и 3-му лицу соответственно, то пропозиции, ассоциированные с вводными словами 2-го лица (*думаешь*, *представляешь*, *знаешь* и т. п.), обычно не принадлежат адресату. Рассматриваются следующие вопросы: есть ли семантическая корреляция между пропозицией и МК, какую иллокутивную функцию имеют МК и пропозиция. Было показано, что некоторые МК употребляются только в вопросительных предложениях.

Ключевые слова: метатекст, вводные слова, адресат, ментальные глаголы

MENTAL PREDICATES IN METATEXT

Kustova G. I. (galinak03@gmail.com)

Vinogradov Russian Language Institute of the Russian Academy of Sciences; Moscow State Pedagogical University; Moscow, Russia

The paper deals with metatext (parenthetical) constructions (MC) with mental verbs (*znat'* 'know', *ponimat'* 'understand', *verit'* 'believe' and the like) in the 2nd person. The following problems are considered: is there a semantic correlation between the proposition and MC; what illocutionary function MC and proposition have. It was shown that some MCs are used only in interrogative sentences.

Key words: metatext, parentheticals, addressee, mental verbs

[1] Исследование выполнено в рамках проекта РНФ № 16-18-02003, осуществляемого в МПГУ. Языковые примеры извлечены из Национального корпуса русского языка, www.ruscorpora.ru.

1. Введение

Метатекстовые (в русской грамматической традиции — вводные) конструкции (далее — МК), как правило, являются результатом редукции и грамматикализации исходных «полноценных» конструкций — предложений или членов предложения: (а) *Посуда бьется к счастью* vs. (б) *К счастью, у него был наш адрес*; (а) *Я надеюсь, что вы не будете возражать* vs. (б) *Вы, я надеюсь, не будете возражать*. При этом один и тот же элемент может иметь разные значения в тексте и в метатексте, ср.: *Пусть будет по-твоему* [‘как ты хочешь’] vs. *По-твоему, я не прав?* [‘ты считаешь’].

Мы будем называть конструкции типа (а) текстовым режимом употребления (для глагольного предиката это означает, что он в качестве матричного глагола присоединяет зависимую клаузу), а употребления типа (б) — метатекстовым режимом. Сравнение с текстовым режимом — важный инструмент анализа МК, и в литературе вводные конструкции часто рассматриваются на фоне текстовых употреблений тех же единиц, ср., например, [Зализняк, Падучева 1987], [Кобозева 1999], [Падучева 1996], [Dehé, Kavalova 2007], [Dehé, Wichmann 2010], [Urmson 1963], [Wierzbicka 1971].

Корпус метатекстовых (вводных) конструкций в русском языке весьма разнообразен как по семантике, так и по способам выражения. В пособиях по синтаксису принято выделять такие группы вводных слов, как модальные (*кажется*, *очевидно*), выражающие логические отношения в тексте (*во-первых*, *следовательно*), эмоциональные (*к счастью*), источник сообщения (*говорят*, *по словам X-а*) и др.

В данной работе мы будем рассматривать МК с ментальными глаголами во 2-м лице: *думаешь*, *знаешь*, *вершь*, *поверь* и под. Эти МК обычно стоят в конце списков вводных компонентов и характеризуются как контактоустанавливающие (фатические). Они используются с целью «привлечь внимание собеседника», «выразить доверительный характер отношений» (*видите ли*, *знаете*, *поверьте*) [Скобликова 2006: 278], в другой формулировке — служат «для подчеркивания, выделения того, что высказывается» [Розенталь 1994: 118]. Но если у них одинаковая функция (да еще и одинаковая форма 2-го лица), можно предположить, что они должны быть синонимичны, как это имеет место в рядах типа *по-моему*, *я думаю*, *по моему мнению*. То есть, теоретически, можно привлечь внимание собеседника с помощью любого предиката — *думаешь*, *знаешь*, *понимаешь* и т. д. Это, разумеется, не так. МК 2-го лица не синонимичны, не взаимозаменимы и употребляются в разных иллокутивных типах предложений, ср.: *А я, думаешь, не испугался?* vs. **А я, знаешь, не испугался?*, но: *А я, знаешь, не испугался*. Следовательно, МК 2-го лица сохраняют лексические значения и выбираются в соответствии с этими значениями, однако в метатекстовом режиме, как будет показано ниже, они функционируют не так, как в текстовом.

Далее принимаются следующие терминологические соглашения: содержание предложения, в котором есть метатекстовый компонент, будем называть ассоциированной пропозицией, или просто пропозицией; метатекстовый компонент 2-го лица будем называть адресатным метатекстовым компонентом, или адресатной метатекстовой конструкцией, сокращенно — АМК.

МК 2-го лица стоят особняком в ряду других МК, как, кстати, и текстовые употребления ментальных глаголов 2-го лица — в ряду других форм ментальных глаголов. Если МК 1-го лица вводят пропозицию говорящего, а МК 3-го лица, соответственно, — пропозицию третьего лица, то МК 2-го лица, вопреки ожиданию, как правило, не вводят пропозицию адресата.

Поясним, что имеется в виду под терминами «пропозиция говорящего» и «пропозиция адресата». Мы будем называть пропозицией говорящего, — как в текстовом, так и в метатекстовом режиме, — то, что вводится ментальным глаголом 1-го лица и является знанием, мнением и т. д. самого говорящего, ср.: *Он, думаю, уже не придет* (*Думаю, что он уже не придет*). Соответственно, пропозиция адресата — это знание, мнение и т. д. адресата.

Очевидно, однако, что адресат не «симметричен» говорящему. Все пропозиции, которые сообщает говорящий в речи, находятся, вообще говоря, в сознании говорящего, и, в широком (тривиальном) смысле, это пропозиции говорящего. Однако когда говорящий передает свои собственные знания, мнения, предположения и т. д., он их, по определению, знает, т. к. передает содержание своего сознания. Если же речь идет о передаче знания, мнения, веры и т. д. другого лица — адресата (или 3-го лица, но этот случай мы не рассматриваем), то говорящий этой информацией исходно не располагает — она находится в сознании адресата.

И действительно, если сконструировать такие повествовательные предложения, где пропозиция (знание или мнение адресата) дается в качестве новой информации (сообщается), ср.: *Ты знаешь, что Петя болеет*; *Ты думаешь, что Петя уехал*, — то такие предложения, будучи грамматически правильными, прагматически неестественны: адресат сам знает, что он думает, а говорящий, напротив, этого не знает, — но даже если и знает, все равно странно сообщать адресату его собственные мнения и знания.

Что же, в таком случае, вводят ментальные глаголы 2-го лица?

Если говорить о текстовом режиме, то наиболее естественным является употребление ментальных предикатов 2-го лица в рамках вопроса: — *Ты думаешь, что от него что-то зависит?*. Если предложение с ментальным глаголом 2-го лица не вопросительное, то пропозиция, вводимая этим глаголом, это либо цитация ранее высказанного адресатом, ср.: *Зря ты думаешь, что X уволится* (адресат сообщил свое предположение ‘X уволится’, поэтому говорящему оно известно); либо актуализация общего знания, ср.: *Вы уже знаете, что он вчера уволился*; либо, чаще всего, реконструкция мыслей адресата: *Вы, наверное, подумали, что он после этого уволился, но это не так*; *Напрасно ты думаешь, что от него что-то зависит*. Для ментальных (и не только ментальных) предикатов 2-го лица характерно употребление в риторических вопросах и других типах экспрессивных высказываний, ср.: *Ты же не думаешь, что он может отказаться!*; *Ты что, не знаешь, как у нас все делается!*; *Ты прекрасно понимаешь, что этого никогда не будет*; *Неужели ты правда веришь, что X уволится* (содержание пропозиции ‘X уволится’ может быть при этом и цитацией, и реконструкцией).

А как ведут себя ментальные глаголы в метатекстовом режиме? Ниже мы рассмотрим три вопроса: 1) согласуются ли пропозиция и вводящий ее ментальный метатекстовый глагол (предикат пропозициональной установки) по типу установки, т. е. содержит ли пропозиция, ассоциированная с АМК *знаешь*, знание, с АМК *думаешь* — мнение, с АМК *веришь* — веру и т. д.; 2) какую иллокутивную функцию имеют АМК и пропозиция (сообщение, вопрос, побуждение); 3) кто контролирует пропозицию, ассоциированную с АМК, — адресат, как это ожидалось бы для глагола 2-го лица, или говорящий.

2. Думаешь, полагаешь, считаешь

Подчиненная пропозиция P при путативных предикатах *думать*, *считать*, *полагать* 1-го или 3-го лица выражает мнение говорящего или мнение 3-го лица соответственно. При метатекстовом употреблении этих глаголов в 1-ом и 3-ем лице картина такая же: *Петя, думаю, на это не согласился бы* — мнение 1-го лица; *А мы, он считает, должны выступить против* — мнение 3-го лица. Напротив, при метатекстовом употреблении глаголов 2-го лица картина другая.

Если в текстовом режиме ментальные глаголы 2-го лица *думаешь* (~*те*), *полагаешь* (~*те*), *считаешь* (~*те*) могут употребляться в повествовательных предложениях (ср.: *Напрасно ты думаешь, что P*), то в метатекстовом режиме эти глаголы употребляются только в вопросительных предложениях. Следовательно, АМК 2-го лица группы *думать* не передают мнение адресата, т. к. говорящий его не знает — иначе он не задавал бы вопрос.

При этом различаются по крайней мере три типа вопросов:

- говорящий действительно хочет узнать мнение адресата — собственно вопрос:

  — *Почему, ты думаешь, он отказался?*

  — *В России, полагаете, будет легче, чем в Японии?*
  [Советский спорт, 2010.01.18];

- говорящий собирается не узнать информацию у адресата (поскольку адресат не знает ответ), а, наоборот, сообщить ему новую информацию — это, так сказать, ложный вопрос, — настоящая цель говорящего — подготовить адресата к тому, что информация будет неожиданной. Для этого типа вопросов характерен глагол *думать*, но не *считать* или *полагать*:

  — *И кому, ты думаешь, дали путевку? — Кому? — Иванову!*;

- говорящий задает полемический (часто — риторический) вопрос, ответ на который очевиден, или реконструирует мнение адресата, чтобы его опровергнуть, показать его ошибочность:

  И начнут тебя там черти на сковородках поджаривать!.. — А тебя, думаешь, не зачнут поджаривать? [М. А. Шолохов. Нахаленок (1925)] = ‘начнут’

А ты, думаешь, одна к нему ездила? [Л. С. Петрушевская. Уроки музыки (1973)] = ‘не одна’

Я даже возмутилась: «Если дали на день рождения, то теперь, ты думаешь, это войдет в систему?» [Александр Терехов. Каменный мост (1997–2008)] = ‘не войдет’

Вы хоть давали себе труд задуматься, откуда брался ксерион? Вы получали его от Фламеля! А Фламель, полагаете, из тумбочки? [Андрей Лазарчук, Михаил Успенский. Посмотри в глаза чудовищ (1996)]

— И никаких следов? — Почему никаких? — Люсин украдкой переглянулся с Гуровым. — На вас, полагаете, мы по наитию вышли? [Еремей Парнов. Александрийская гемма (1990)]

Вот этот ваш Кривченя… он просто так, считаете, стелился перед немцем, выпрашивая вам свободу [Сергей Самсонов. Одиннадцать (2010)].

3. Представляешь

В текстовом режиме *представляешь* (~*те*) может функционировать как повествовательное: *Ну, ты представляешь, сколько у нас чиновники получают* — хотя это, строго говоря, не сообщение, а констатация. Более типичным является вопрос. Здесь можно выделить три типа вопросов.

Первый тип — обычный вопрос: *Ты хотя бы примерно представляешь, куда нам теперь идти?*

Второй тип — предваряющий, контактоустанавливающий вопрос, который должен подготовить адресата к тому, что он сейчас услышит какую-то интересную, удивительную или неожиданную информацию: *Представляешь, кого я сейчас встретил?*

Третий тип — аналог риторического вопроса, не требующего ответа: *Представляешь, сколько у нас чиновники получают!* = ‘много’; *Представляешь, как я испугался!* = ‘очень’. Это не классический риторический вопрос — в смысле экспрессивного отрицания. Но все равно это экспрессивное высказывание, которое не запрашивает информацию, а сообщает ее.

Итак, само предложение с текстовым *представляешь*, как правило, вопросительное, и в нем встречаются те же три типа вопросов, которые мы рассмотрели в предыдущем разделе. Т.е. текстовое *представляешь* похоже на метатекстовое *думаешь*. Напротив, метатекстовое *представляешь* значительно отличается от текстового *представляешь*.

АМК *представляешь* бывает только вопросительным: *Завтра, представляешь, занятия отменяют!* Хотя в русской пунктуации не принято ставить отдельный знак препинания во вводных конструкциях (в работе [Кобозева 1999: 542] для таких конструкций используется знак «?»), *представляешь* произносится с вопросительной интонацией (утвердительное *представляешь* бывает не во вводной, а в так называемой вставной конструкции: *А эта деревушка — ну,*

ты представляешь — находится высоко в горах). Хотя в метатексте *представляешь* всегда вопросительное, оно, в отличие от метатекстового *думаешь* и в отличие от текстового *представляешь*, несовместимо с вопросом — ни с прямым, ни с риторическим. Можно сказать: *И кто, вы думаете, это был?*; *А я, думаешь, этого не хочу?*, но невозможно: **И кто, представляете, это был?*; **А я, представляете, этого не хочу?*

Т. е. предложение, в которое входит вопросительное вводное *представляешь*, само никогда не бывает вопросительным, а является повествовательным либо восклицательным:

> *Главное, представляешь, говорит: ой, Марина, да что вы все с этими деньгами!* [Андрей Волос. Недвижимость (2000) // «Новый Мир», 2001]

> *Вчера, представляете, жвачку купил. Пожевал жвачку и вдруг проглотил* [И. Э. Кио. Иллюзии без иллюзий (1995–1999)]

(здесь удивительный факт сообщается не в этом предложении, а в следующем — *проглотил*).

> *Сказали, что нельзя приходить после девяти вечера, представляешь?* [«Столица»];

> *Он плохо знал, допустим, математику, а указывал, и, представляете, — сходилось* [Д. Гранин. (1987)].

Т. е. метатекстовое *представляешь* соотносится с тем типом текстового употребления *представляешь*, когда говорящий хочет не получить информацию, а сообщить ее (ср.: *Представляешь, кого я сейчас встретил?*), и маркирует эту информацию как необычную, неожиданную, как бы призывая адресата разделить это удивление (варианты: *Можешь себе представить?*; *Ты только представь!*). Отличие в том, что информация уже сообщена (*Так спешил и, представляешь, все-таки не успел*).

Поскольку в таких предложениях сообщается неизвестная адресату информация, было бы логично использовать форму *представь / представьте* — и дальше эту информацию сообщить.

Разумеется, форма императива тоже функционирует в метатекстовых конструкциях (*А я, представьте, так и думал*), но эквивалентом *представляешь* она, как правило, не является, а выражает несколько другой смысл: *Он плохо знал, допустим, математику, а указывал, и, представьте, — сходилось*. Здесь тоже есть идея сообщения неожиданного факта, но она сочетается с идеей разоблачения неверного представления адресата: ‘вы, наверное, подумали, что не сходилось (раз он плохо знал математику), но вы ошиблись’.

Наконец, есть, условно говоря, «полемическое» (некооперативное, конфликтное) *представь* (чаще — *представь себе*), оно употребляется в самых разных речевых актах (упрек, возражение, опровержение): *Вы только начинаете, а я, представьте, уже все закончил!*

4. Веришь / не поверишь

Веришь похоже на *представляешь*: функционирует как вопрос, но вставляется в сообщение и содержит новую для адресата информацию.

4.1. Веришь

Отоварил одного… Потом, веришь, полночи не спал, жалко было… [Сергей Каледин. Записки гробокопателя (1987–1999)]

— Я, веришь, коммуняк зрить не могу, — сказал Синяк. [Сергей Каледин. Записки гробокопателя (1987–1999)]

4.2. Веришь ли

Павлу Петровичу, государю нашему бесценному, — веришь ли? — видение было. Привиделось ему: сорока шести лет ему не пережить [Борис Евсеев. Евстигней // «Октябрь», 2010]

Я знаю, что есть какие-то Деникин и Колчак, и что они хотели взять Питер и чуть было не взяли его, — но, веришь ли, мое решение никак с ними не связано [Дмитрий Быков. Орфография (2002)]

4.3. Верите

— Иной раз посмотришь на человечков и, верите, только диву даешься: при этакой энергии — целы! [Леонид Зорин. Глас народа (2007–2008)]

4.4. Верите ли

Сто двадцать одну Маргариту обнаружили мы в Москве, и, верите ли, — тут Коровьев с отчаянием хлопнул себя по ляжке, — ни одна не подходит [М. А. Булгаков. Мастер и Маргарита (1929–1940)]

Водочки в трактире выпил, и, верите ли, ни в одном глазу, весь хмель в обиде сгорает… [Леонид Юзефович. Костюм Арлекина (2001)]

Здесь адресату сообщается факт, который говорящему кажется странным, не соответствующим обычному опыту, привычным представлениям или контексту. Поскольку речь идет о факте, неудивительно, что данное употребление *веришь* похоже на один из типов употребления АМК *знаешь* (см. ниже), ср.: *И, веришь, мне тогда было все равно — И, знаешь, мне тогда было все равно*, — хотя в текстовом режиме эти глаголы выражают, в каком-то смысле, противоположные пропозициональные установки.

В то же время, хотя речь идет о факте, употребление глагола *веришь* здесь оправданно в том смысле, что этот факт, вообще говоря, адресат знает только со слов говорящего и не может проверить. Кроме того, поскольку говорящий оценивает этот факт как странный, необычный, нетривиальный, он как бы

допускает, что адресат мог бы и не поверить, — именно это допущение и маркирует АМК *веришь / верите*. При этом говорящий, конечно, не предполагает, что адресат ему не верит: *веришь* выполняет функцию оговорки, смягчения утверждения. Тем самым вводное *веришь*, в отличие от *представляешь*, совершенно не соотносится с текстовым употреблением, т. е., например, предложение *Потом, веришь, полночи не спал* не является преобразованием предложения: *Ты веришь, что я полночи не спал?*

Существенно иначе ведет себя, казалось бы, синонимичный АМК *не поверишь / не поверите* — здесь говорящий тоже сообщает необычный или удивительный факт, но *не поверишь* не вопросительное, а повествовательное, поскольку это гипотеза говорящего (‘думаю, что ты можешь не поверить’):

> *— Люсинда, ты где эту кофточку отхватила? —*
> *В «Эскаде», не поверишь, всего за восемьсот бакинских!*
> [Наталья Александрова. Последний ученик да Винчи (2010)]

> *На Филиппинах, не поверишь, по ночам у меня зуб на зуб не попадал, надевал всю одежду, какая была* [«Русский репортер», 2014]

> *А я тебя, не поверишь, только сегодня вспоминала*
> [Маша Трауб. Бочка (2009)]

> *Кто-то читает «Playboy», а кто-то, не поверишь, журнал про ножи «ПРОРЕЗ»!* [«Хулиган», 2004.06.15].

Употребления *веришь*, *не поверишь* связаны с речевыми формулами *Ты не поверишь!*; *Ты можешь в это поверить?* (последнее выражение может употребляться и в качестве МК: *А он, ты можешь в это поверить, взял и уволился*).

5. Знаешь

В текстовом режиме в повествовательных предложениях *знаешь* прагматически странно (странно сообщать адресату то, что он знает: *?Ты знаешь, что завтра собрание*; для этого нужны специальные условия, например, полемический контекст: *Ты же знаешь, что завтра собрание!*). Текстовое *знаешь* встречается, в основном, в вопросах, причем оно имеет особый тип употребления, который упоминает Ю. Д. Апресян: *Знаешь, кто на тебя нажаловался?* Это может быть и нормальный вопрос, на который говорящий хочет получить ответ, но в [Апресян 1995] имеется в виду особое значение — в таких предложениях предполагается, что у адресата нет соответствующей информации, а у говорящего она есть, и он собирается ее сообщить.

В метатекстовом режиме *знаешь*, так же как *представляешь* и *веришь*, не сочетается с вопросительным предложением, а сочетается с сообщением, при этом *знаешь* имеет «ослабленное значение „доверительности“» [Апресян 1995: 424]:

> *— Спасибо, Ладушка, но мне, знаешь, не до твоего чая*
> [В. П. Катаев. Алмазный мой венец (1975–1977)];

А я вот, знаете, приду ещё затемно, сяду на этот вот камешек — я его специально со склона скатил — и сижу, сижу [Ю. О. Домбровский. Факультет ненужных вещей (1978)];

— Какой-то, знаете, он у меня невезучий. На других посмотришь — прямо в руки всё идёт [Андрей Волос. Недвижимость (2000)].

Идея данного употребления *знаешь* такая же, как и у *представляешь*: адресат не знает, а говорящий ему сообщает. Но какова, в таком случае, функция АМК *знаешь*?

В случае *Знаешь, кто на тебя нажаловался?* исходное значение *знаешь* хоть и в ослабленном виде, но сохраняется: говорящий как бы спрашивает: Ты знаешь Р? (думаю, что не знаешь, но на всякий случай интересуюсь); если не знаешь, могу тебе сообщить.

Поскольку в метатекстовом режиме *знаешь* с вопросом не сочетается, то можно было бы предположить, что происходит полная десемантизация — этот показатель используется как в чистом виде фатический (ср.: *Слушай, сейчас такое было…*), — чтобы просто привлечь внимание собеседника. Однако это не так. *Знаешь* имеет разную интерпретацию в зависимости от смысла основного высказывания, и, вдобавок, произносится с разной интонацией.

Совсем редуцированное, так сказать, безударное *знаешь / знаете* апеллирует к некоторому обобщенному опыту, который может быть у разных людей, в том числе и у адресата (‘вы, наверное, знаете, как это бывает’), или к воображению адресата:

Бизнес есть бизнес. Там, знаешь, не поспишь [А. Волос. Недвижимость (2000)]

А у этих, посетителей, знаете, у кого мать и отца расстреляли, кто по всяким местам промыкался [«Экран и сцена», 2004.05.06].

Такое *знаешь* может произноситься и с вопросительной интонацией:

А на мне были ботиночки такие новые красивые, знаешь, такие коричневые, высоко зашнурованные, как тогда носили, и они мне были немножко малы, так что ноги замерзли [Марина Палей. Поминовение (1987)] —

говорящий как бы исходит из того, что адресат может что-то знать о сообщаемом, но сообщает все-таки свое собственное знание.

Наконец, говорящий может апеллировать не к обобщенному, а к личному опыту, знанию адресата. И здесь *знаешь* похоже на вопрос-напоминание, вопрос-актуализацию — и по интонации, и по функции:

Она там с одним фертом ходила, знаешь, из этаких, из свободных художников [Ю. О. Домбровский. Факультет ненужных вещей (1978)] —

говорящий предполагает, что адресат видел и может представить себе таких свободных художников, т. е. он хочет актуализировать этот опыт.

Бывает экспрессивное *знаешь* (с особой экспрессивной интонацией) — например, если говорящий сообщает что-то нетривиальное, интересное или

просто подчеркивает важную для него мысль, на которую должен обратить внимание и адресат:

> *Машину собирается покупать. Вообще, знаешь, настоящая европейская женщина! Сколько она меня мучила за последние годы, а я всё-таки восхищаюсь!* [Ирина Муравьева. Мещанин во дворянстве (1994)],

ср. текстовое: *Знаешь, как я испугался!*.

Бывает «полемическое» *знаешь*, когда говорящий как бы возражает на предыдущую реплику или на какую-то ситуацию:

> *«Ну, знаете… можно и зайца научить курить…»* [Форум (2006–2010)]

> *Пугают какими-то восемьюдесятью пятью процентами, которые якобы «поддержали арест мошенника и кровопийцы Гусинского»? — Да ну, погодите, это ж факт, что абсолютное большинство народа с удовольствием эти меры поддержит! — Ну, знаете, эта часть народа много ещё чего хорошего поддержит* [«Коммерсантъ-Власть», 2000].

Возможно, это и есть переходное звено от исходного *знаешь* к полностью редуцированному, фатическому. Но связь с фактивностью даже в таком редуцированном употреблении, когда предикат превращается практически в частицу, все-таки сохраняется: *знаешь* вводит именно достоверную информацию.

Итак, рассмотренный материал показывает, что если МК 1-го лица (*думаю*, *знаю*, *верю*) вводят пропозицию говорящего (что естественно), то МК 2-го лица во многих случаях не вводят пропозицию (знание, мнение) адресата, как это ожидалось бы от глагола 2-го лица, а тоже вводят пропозицию говорящего.

За исключением *думаешь*, которое вставляется в вопрос — обычный, когда говорящий хочет узнать мнение адресата (*Он, думаешь, согласится?*), или риторический, когда говорящий уже знает (или реконструировал) мнение адресата и полемизирует с ним (*Руководство, думаете, поддержит такой безумный план?!*), — все остальные рассмотренные вводные глаголы 2-го лица вставляются в повествовательное или восклицательное (экспрессивное) предложение. При этом, хотя большинство этих глаголов путативные, содержащее их предложение передает не мнение, а знание говорящего, достоверную информацию, которая часто является новой для адресата. Это говорит о том, что при введении метатекстовых ментальных глаголов 2-го лица в предложение действуют совершенно другие механизмы, чем при введении глаголов 1-го лица (конструкции с глаголами 1-го лица являются, как правило, просто редукцией исходной текстовой структуры, ср.: *Я думаю, что он уехал — Он, я думаю, уехал*).

Вопросительная интонация, с которой произносятся большинство вводных ментальных глаголов 2-го лица (кстати, тот факт, что иллокутивная функция вводного глагола (вопрос) не совпадает с иллокутивной функцией всего предложения (сообщение), является лишним доказательством того, что АМК — особый тип вводных конструкций), действительно, является средством привлечения внимания адресата. Однако, вопреки распространенному мнению (см., например, [Скобликова 2006]), вводные глаголы 2-го лица в большинстве случаев не являются просто контактоустанавливающими, фатическими

(за исключением слов-паразитов типа *понимаешь*, ср.: *А я, понимаешь, ничего, понимаешь, не нашел, понимаешь*), а служат для того, чтобы передать отношение говорящего к сообщаемому как к чему-то необычному, труднопредставимому, удивительному, неожиданному (*Он, представляешь, оказался сыном министра*); или, наоборот, как к чему-то общепонятному или легкопредставимому (*Вообще она, знаешь, настоящая европейская женщина!*); или же как к чему-то важному для самого говорящего, о чем должен знать и адресат (*Какой-то, знаете, он у меня невезучий*), — и имеют самостоятельную семантику, о чем свидетельствует тот факт, что АМК с ментальными глаголами не являются взаимозаменимыми.

Литература

1. *Apresyan Yu. D.* (1995), The problem of factivity: to know and synonyms [Problema faktivnosti: znat' i ego sinonimy] // Apresyan Yu. D. Selected Works, v.2 [Izbrannye trudy, t. 2]. Moscow, YaSK, pp. 403–433.
2. *Zaliznyak A. A., Paducheva E. V.* (1987), On the semantics of parenthetical verbs [O semantike vvodnogo upotrebleniya glagolov] // Problems of cybernetics [Voprosy kibernetiki]. Moscow, pp. 80–96.
3. *Kobozeva I. M.* (1999), On two types of constructions with a parenthetical verb [O dvukh tipakh vvodnykh konstruktsii s parenteticheskim glagolom] // E. V. Rakhilina, Ya. G. Testelets (red.). Typology and theory of language: from description to explanation [Tipologiya i teoriya yazyka: ot opisaniya k ob"yasneniyu. K 60-letiyu A. E. Kibrika]. Moscow, pp. 539–543.
4. *Paducheva E. V.* (1996), Semantic researches [Semanticheskie issledovaniya], Moscow, «Yazyki russkoi kul'tury».
5. *Rozental D. I.* (1994) Handbook of spelling, pronunciation, literary editing [Spravochnik po pravopisaniyu, proiznosheniyu, literaturnomu redaktirovaniyu]. Moscow.
6. *Skoblikova E. S.* (2006), Modern Russian: Syntax of a simple sentence [Sovremennyi russkii yazyk: Sintaksis prostogo predlozheniya]. Moscow, «Flinta»-«Nauka».
7. *Dehé N., Kavalova Y.* (2007) Parentheticals. (Linguistik Aktuell/Linguistics Today 106) John Benjamins,.
8. *Dehé N., Wichmann A.* (2010) Sentence-initial *I think (that)* and *I believe (that)*. Prosodic evidence for use as main clause, comment clause and discourse marker. Studies in language. Vol. 34. No 1. Pp. 36–74.
9. *Urmson J. O.* (1963), Parenthetical verbs // Ch. E. Caton (ed.) Philosophy and ordinary language. Urbana, Chicago, London: University of Illinois Press, pp. 220–240.
10. *Wierzbicka A.* (1971), Metatekst w tekscie // O spójności tekstu. Wrocław-Warszawa, pp. 105–121.

RUSSIAN WORD SENSE INDUCTION BY CLUSTERING AVERAGED WORD EMBEDDINGS

Kutuzov A. B. (andreku@ifi.uio.no)
University of Oslo, Oslo, Norway

The paper reports our participation in the shared task on word sense induction and disambiguation for the Russian language (RUSSE'2018). Our team was ranked 2nd for the wiki-wiki dataset (containing mostly homonyms) and 3rd for the bts-rnc and active-dict datasets (containing mostly polysemous words) among all 19 participants.

The method we employed was extremely naive. It implied representing contexts of ambiguous words as averaged word embedding vectors, using off-the-shelf pre-trained distributional models. Then, these vector representations were clustered with mainstream clustering techniques, thus producing the groups corresponding to the ambiguous word' senses. As a side result, we show that word embedding models trained on small but balanced corpora can be superior to those trained on large but noisy data—not only in intrinsic evaluation, but also in downstream tasks like word sense induction.

Keywords: lexical semantics, word sense induction, word sense disambiguation, word embeddings, distributional semantics, clustering

ИЗВЛЕЧЕНИЕ ЛЕКСИЧЕСКИХ СМЫСЛОВ В РУССКОМ ЯЗЫКЕ НА ОСНОВЕ КЛАСТЕРИЗАЦИИ УСРЕДНЕННЫХ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ СЛОВ

Кутузов А. Б. (andreku@ifi.uio.no)
Университет Осло, Осло, Норвегия

В этой статье мы описываем наше участие в соревновании по извлечению лексических смыслов неоднозначных слов в русском языке (RUSSE'2018). Среди 19 участников наша команда заняла второе место при оценке на датасете wiki-wiki (состоящем в основном из омонимов) и 3 место при оценке на датасетах bts-rnc и active-dict (состоящих в основном из полисемичных слов со связанными значениями). Мы намеренно использовали чрезвычайно простой метод решения задачи. Он состоял

в представлении контекстов неоднозначных слов в виде усредненных векторов составляющих их лемм, взятых из готовых предобученных дистрибутивных моделей. Затем эти векторные репрезентации кластеризовывались (стандартными алгоритмами) на группы, соответствующие значениям неоднозначного слова. В качестве побочного результата, мы показываем, что для целей обучения дистрибутивных моделей сравнительно небольшие, но сбалансированные корпусы могут превосходить по качеству огромные, но зашумленные и несбалансированные текстовые массивы—не только при оценке на искусственных датасетах, но и в таких практических задачах, как извлечение лексических смыслов.

Ключевые слова: лексическая семантика, извлечение смыслов, разрешение лексической многозначности, дистрибутивная семантика, кластеризация

1. Introducing word sense induction task

Human language is inherently ambiguous on all of its tiers. Grammatical and syntactic ambiguity is successfully solved by part-of-speech taggers and dependency parsers. But this is not enough, as morphologically and syntactically identical words can possess different senses or meanings. Indeed, all that happens with semantics, happens at the level of word senses, not words. This means that some ways of disambiguating ambiguous words and finding out the correct number of senses have to be devised.

Word sense induction (WSI) is an important part of computational lexical semantics and boils down to the task of automatically discovering the senses of semantically ambiguous words from unannotated text. It has long research history for English and other languages, with several relevant *SemEval* shared tasks [14]. However, until recently, the NLP community lacked proper evaluation of WSI methods for Russian. RUSSE'2018 shared task [16][1] fills in this gap. This paper describes the approach we used in the framework of this competition.

The participants of the shared task were given three sets of Russian utterances containing semantically ambiguous words. The participating systems had to group the utterances containing a particular ambiguous word into clusters, depending on the sense this word takes in this particular utterance.

The organizers offered two tracks:

1. *Knowledge-rich*, where the participants were permitted to use dictionaries or other lexical databases containing sense inventories;
2. *Knowledge-free*, where participants were allowed to use only text corpora and models automatically derived from these corpora.

We participated in the *knowledge-free* track. Thus, we had to infer word senses from the data, without relying on any external sources like *WordNet* [13], *BabelNet* [15] or *Wiktionary*[2]. The performance of the systems was evaluated by calculating the

[1] https://russe.nlpub.org/2018/wsi/

[2] https://ru.wiktionary.org

Adjusted Rand Index (ARI) between context clustering produced by the systems and the gold clustering provided by the organizers.

We intentionally employed a very simplistic (even naive) approach to word sense induction, which we describe below. The reason for this was that we were interested in whether Russian WSI task can be solved using only already available algorithms and off-the-shelf models. It turned out to be true for one of the three RUSSE'18 datasets (we ranked 2nd) but not so true for other two (we ranked 5th). It should be noted, however, that none of the participants achieved reasonably high scores for these last 2 datasets. We outline the differences between the datasets in Section 3.

Overall, our contributions are twofold:

1. We describe and publish the WSI system for Russian, which produces very competitive results for homonyms with non-related senses, and which is based exclusively on off-the-shelf tools and models.
2. It was already known that training corpus balance can be even more important for word embedding models than its size, when evaluated intrinsically. In this paper, we show that this holds for extrinsic evaluation setting as well, with WSI as a downstream task in this case.

The rest of the paper is organized as follows. Section 2 briefly outlines the previous work related to word sense induction and distributional semantics. In Section 3, we present the datasets offered by the shared task organizers and the corpora used to train our word embedding models. Section 4 provides the details of the employed approach. In Section 5 we describe the results, comparing them to other participants, and in Section 6 we conclude.

2. Related Work

Word sense induction task is closely related to word sense disambiguation: the task to assign meanings to ambiguous words from a pre-defined sense inventory. Even this easier task is notoriously difficult to handle computationally. In 1964, Yehoshua Bar-Hillel, Israeli mathematician and linguist even proclaimed that "'sense ambiguity could not be resolved by electronic computer either current or imaginable" [1].

Fortunately, it turned out that things are not that bad. Since the sixties, many word sense disambiguation techniques appeared, which were quite successful in telling which sense the particular word is used in. In the recent years, the majority of these techniques are based on statistical approaches and machine learning.

However, all word sense disambiguation approaches suffer from the same problem known as knowledge acquisition bottleneck. They need ready-made sense inventory for each ambiguous word: otherwise, there is nothing to choose from. Manually annotated semantic concordances and lexical databases quickly get outdated. They don't keep up with the changes in language, and humans simply cannot annotate that fast. This is especially true for named entities and for specialized domains.

At the same time, it is relatively easy to compile large up-to-date corpora of unannotated text. It is then possible to infer word sense inventories from these corpora automatically. This task is called *unsupervised word sense disambiguation* or *word sense*

induction (WSI): the input is corpus, and the output consists of sense sets for each content word in the corpus we are interested in. Quoting Adam Kilgariff in [6], "*word senses are abstractions from clusters of corpus citations*".

Thus, there are no pre-defined sense inventories: we discover senses for a given word directly from text data. This boils down to the task of clustering occurrences of the input word in the corpus, based on their senses.

The foundations for clustering-based WSI were laid in [5] and [19]. In its essence, it is a very straightforward approach based on word distributions:

1. Represent each ambiguous word with a list of its context vectors;
   - context vector contains identifiers of context words in a particular context (sentence, phrase, document, etc...).
2. For each word, cluster its lists into a (predefined) number of groups, using any preferred clustering method;
3. For each cluster, find its centroid;
4. These centroids serve as sense vectors for the subsequent word sense disambiguation.

At test time, the system is given a new context (for example, sentence) containing an ambiguous input word. It computes its context vector by listing the context words, and then chooses the sense vector which is the most similar to the current context vector.

Of course, by the nature of the approach, the induced "senses" are coarse, nameless and often not directly interpretable (see [17] for an attempt to overcome non-interpretability). However, it is still possible to tell one sense from another in context, and this is what real-world systems need. Further on, the WSI approaches were enriched with additional techniques, for example with lexical substitution [22]. Today, WSI is extensively relied upon in many NLP tasks, including machine translation and information retrieval [14].

We use prediction-based word embedding models of lexical semantics as the source of distributional information representing word meanings. This sort of models is extensively described elsewhere. See [12] and [2] for the background of *Continuous Skipgram* and *fastText* algorithms that we employed.

Note that there are many other WSI algorithms, including graph-based approaches. We refer the interested reader to [3] for the general overview and to [10] for an example of the application of graph-based WSI for Russian data. Very recent experiments with combining graph and word embedding approaches to WSI are described in [23].

3. Data overview

In this section, we describe the RUSSE'18 datasets, and the word embedding models we used to process them.

RUSSE'18 shared task offered three datasets (with a training and a test part in each):

1. **wiki-wiki:** sense inventories and contexts from the Russian Wikipedia articles
2. **bts-rnc**: sense inventories from "*Bolshoi Tolkovii Slovar*" dictionary (BTS), contexts from the Russian National Corpus [15]

3. **active-dict**: sense inventories from the *Active Dictionary of the Russian Language*, contexts from the examples in the same dictionary.

Each training set consisted of several ambiguous query words (from 4 in the **wiki-wiki** to 85 in the **active-dict**) and about a hundred contexts for each of them. The context as a rule included several sentences, not more than 500–600 characters total. Each context was annotated with the identifier of the sense in which the corresponding query word was used in this context. The test sets featured the same structure, of course without the sense annotation. Thus, the task was to find out for each query word in the test set how many senses it has and which contexts belong to the same senses.

The systems' performance for each dataset was evaluated separately. We strongly support this decision of the organizers and argue that it might even make sense to cast this as two independent shared tasks.

The reason is that **wiki-wiki** dataset is substantially different from the other two. First, its sense structure is much more stable: the training set query words have exactly two senses each. At the same time, for the **bts-rnc** training set the average number of senses per query word is 3.2, and the maximum number of senses is as high as 8. The **active-dict** training set is even more varied, with the average number of senses 3.7, and the maximum number of senses 17 (*sic!*).

As if this was not enough, the nature of these senses is unsurprisingly different. In the **wiki-wiki** dataset, most senses are homonyms, that is unrelated to each other (for example, "бор" *pine wood* and "бор" *boron*). On the contrary, the other two datasets are abundant in polysemy, where word senses are somehow related. Cf. "обед" *lunch* and "обед" *lunchtime* from the **bts-rnc** dataset, or "дерево" *tree* and "дерево" *wood* from the **active-dict** dataset. There are also many cases of metonymy and other subtle semantic shifts.

Of course, word senses are a kind of continuum, and there is no distinct boundary between homonymy and polysemy. Even for human experts, it is often difficult to tell how many senses does a word really have. However, we still think that the **wiki-wiki** dataset presents a very different task. This task (*inducing meanings of homonyms*) is much easier than the task of *inducing different senses of polysemous words*. Arguably, considerably different approaches are needed for both.

Anyway, to handle semantic phenomena, one needs a way to model semantic similarities and dissimilarities between words. To this end, we employed pre-trained word embedding models for Russian, downloaded from the *RusVectōrēs*[3] web service [8]. We tested five models:

1. **ruscorpora_upos_skipgram_300_5_2018** trained on the Russian National Corpus [18] (about 250 million words);
2. **ruwikiruscorpora_upos_skipgram_300_2_2018** trained on concatenation of the Russian National Corpus and the Russian Wikipedia (about 600 million words);
3. **news_upos_cbow_600_2_2018** trained on a large Russian news corpus (about 5 billion words);
4. **araneum_upos_skipgram_300_2_2018** trained on the Araneum Russicum Maximum web corpus [26] (about 10 billion words);

[3] http://rusvectores.org/ru/models/

5. **araneum_none_fasttextskipgram_300_5_2018** trained on the same corpus as the previous model, but using the fastText algorithm instead of the Continuous Skipgram.

With these components at hand, we attempted to build a system capable of inducing word senses for the three datasets. In the next section, we describe this system.

4. Our approach

We applied more or less the same workflow for all the three datasets, with minor alterations, depending on what worked best. Briefly, our approach can be summarized in the following steps:

1. Lemmatize and PoS-tag contexts;
2. Represent each context as a fixed-length vector manifesting its semantics;
3. Determine the number of clusters in the set of contexts, using the *Affinity Propagation* algorithm;
4. Group the contexts into clusters representing word senses, using either the same *Affinity Propagation* or other clustering algorithm.

There are two important and practically independent phases in this workflow, which we describe in the next 2 subsections.

4.1. Contexts representations

The first phase consists of converting context utterances from lists of words to fixed length vector representations. Note that first we lemmatized and PoS-tagged all words in the context utterances using *UDPipe* 1.2 tagger [21] trained on Russian Universal Dependencies corpus [4]. We also tried to use *Mystem* tagger [20] instead, but this did not result in any improvements for the WSI task. The ambiguous query words themselves were removed from the utterances.

Then, for each lemmatized context utterance, we created "semantic fingerprints" as described in [7]. The "fingerprint" function takes as an input the list of lemmas and a pre-trained word embedding model. It looks up the embeddings for all the lemmas from the context utterance present in the model's vocabulary. Then, these vectors are averaged to produce the function output, which is a single vector of the same dimensionality as the vectors in the employed model (we used the models with the vector size 300). This dense vector is used as a semantic representation of the context utterance.

Note that we slightly modified the "semantic fingerprint" notion from [7]. First, we counted multiple occurrences of the same lemma as one occurrence (that is, binary bag-of-words was used, discarding local word frequencies in the context utterances). Second, before averaging the word vectors, we assigned them weights in the range of [0...1], in inverted proportion to the word frequencies in the training corpus of the underlying word embedding model. This way, "globally frequent" words (which are often not sense-specific) got less influence on the resulting semantic fingerprints,

while "globally rare" words (often specific for a particular sense) became more influential. In our experience, both changes improved the word sense induction performance (see Section 5).

With the vector representations of contexts ("semantic fingerprints") ready, it is possible to cluster them into groups corresponding to different senses of the query word.

4.2. Contexts clustering

Theoretically, any clustering algorithm can be used in this case. The only complication is that the number of senses (and thus the number of clusters) for any given query word is unknown. It means that this number must be induced from the data.

Many clustering techniques are able to do this. We employed the *Affinity Propagation* algorithm: first, because it is readily accessible in the *scikit-learn* library[4], and second, because it was successfully applied to related tasks (in [22] for English and in [9] for Russian).

Affinity Propagation produces clustering of the contexts, which can be used immediately as the desired sense-specific grouping. For the **wiki-wiki** dataset, this was our strategy. However, for two other datasets, we found that our system performs better if we use *Affinity Propagation* only to induce the number of clusters (senses). After that, another clustering algorithm (either *K-Means* or spectral clustering) is called to separate the data into the induced number of groups. This once again emphasizes the differences between the datasets in the shared task.

Note that the *Affinity Propagation* takes two parameters: preference and damping, which both greatly influence the behavior of the algorithm, especially the resulting number of clusters. We performed grid search to determine the best combination of these parameters for each dataset[5].

[4] We also tested DBSCAN clustering algorithm, but it yielded suboptimal results for all datasets.

[5] The best values for the preference parameter seem to lie between -0.6 and -0.7, while for the damping parameter the sweet spot is between 0.7 and 0.8.

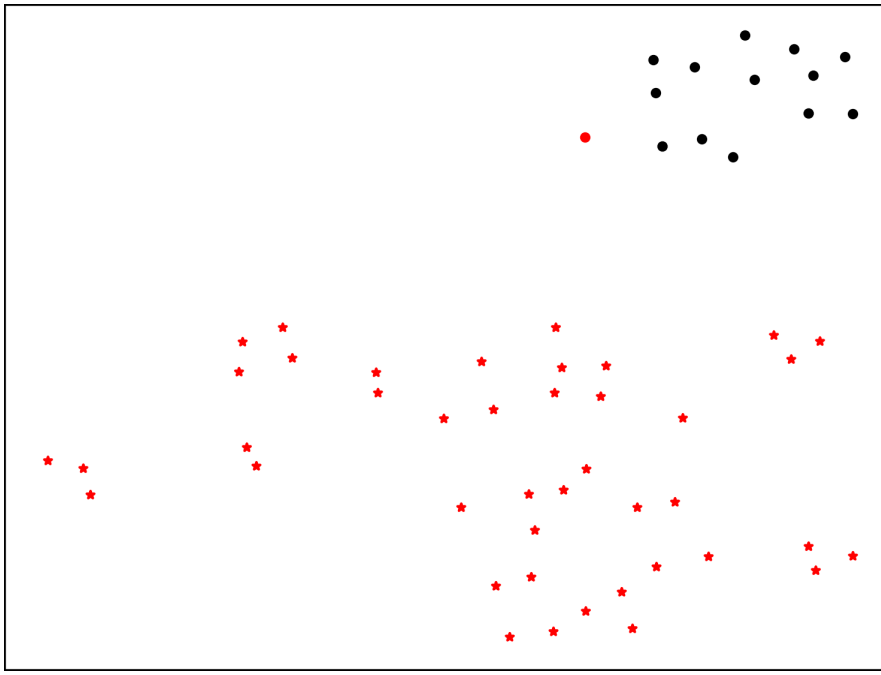


Figure 1: Clustering of the «бор» contexts ("pine wood" and "Boron"). Colors are clusters assigned by the system, shapes are gold clusters.

The resulting system, despite its simplicity, produces reasonable clusterings. We illustrate this with the Figure 1 which presents the 2-dimensional *t-SNE* projection of 300-dimensional context vector representations for the query word "*бор*" mentioned above. Stars stand for the contexts annotated with the "*pine wood*" sense, and circles for the "*Boron*" sense. Colors reflect the clustering produced by the system. One can see that it successfully detected the correct number of clusters (2) and correctly grouped all the contexts, except one.

5. Results

We first present the results of our experiments on the training data, and then describe the performance of the presented system on the test sets in comparison with other participants of the shared task.

As mentioned before, we experimented with five pre-trained word embedding models. The Table 1 provides an overview of the best results that we got for each dataset using each particular model as the source of knowledge about word meanings.

It is clear that the model trained exclusively on the Russian National Corpus (RNC) was the best for all three datasets, despite comparatively small size of the corpus. This further supports the importance of proper compiling and balancing the training corpora for word embedding models. It was previously shown in [11] that the

models trained on the RNC are very often not worse or even better than those trained on a much larger web corpus in the intrinsic evaluation (semantic similarity task). The present work continues this line of research and proves that this holds at least for some extrinsic evaluation settings as well (WSI in this case).

The way word vectors are averaged to produce "semantic fingerprints" greatly influences the results for the **wiki-wiki** dataset, as shown in Table 2. Changing the representation to binary bag-of-words instead of count bag-of-words brings stable improvements, as well as introducing global frequency weights. The other 2 datasets are almost agnostic to these parameters: as we believe, precisely because of their different nature. Note also that due to the usage of the second clustering algorithm (dependent on random initialization), the results for the **bts-rnc** and **active-dict** datasets are non-deterministic and fluctuate slightly from one run to another.

Table 1: Clustering performance (ARI) on the training sets, depending on the pre-trained word embedding model

Model / Dataset	*wiki-wiki*	*bts-rnc*	*active-dict*
ruscorpora_upos_skipgram_300_5_2018	**0.772**	**0.176**	**0.260**
ruwikiruscorpora_upos_skipgram_300_2_2018	0.669	0.162	0.210
news_upos_cbow_600_2_2018	0.653	0.174	0.143
araneum_upos_skipgram_300_2_2018	0.492	0.162	0.197
araneum_none_fasttextskipgram_300_5_2018	0.695	0.171	0.178

Table 2: Clustering performance (ARI) depending on the parameters of word vector averaging

Dataset	**Original semantic fingerprints**	**+ binary bag-of-words** (discarding local word frequencies)	**+ weights** (global word frequencies)
wiki-wiki	0.579	0.717	**0.772**
bts-rnc	0.169	0.167	0.176
active-dict	0.250	0.254	0.260

Finally, the Table 3 presents our scores on the test sets, and thus, the resulting performance of the presented system. To cut it short, our naive approach turned out to be very competitive for the WSI on homonyms from the **wiki-wiki** dataset, winning the 2nd place in the ranking with the ARI of 0.71.

For more subtle inter-related senses of the **bts-rnc** and **active-dict** datasets, our approach performed much worse, although still allowing us to stay in the top 25% results. Note that for these two datasets, none of the competing systems managed to achieve ARI higher than 0.34, which is a long way to any possible production usage. Partly this may be caused by flaws in the gold data itself: it would be an interesting research to measure human performance and inter-rater reliability in clustering contexts for these two datasets. It is quite probable that it will turn out to be not much higher.

It is also interesting that the best results for the **wiki-wiki** (including ours) and **bts-rnc** datasets outperform state-of-the-art WSI results for English, which achieve ARI about 0.215–0.286 [24, 25]. Certainly, this can be caused by the differences between the RUSSE'18 datasets and those of SemEval-2013 and WWSI, but still this phenomenon deserves a deeper analysis in the future.

Table 3: Overall shared task results (evaluated on the test sets)

	Our ARI ("RusVectores" team)	Rank (of 19 participants)	The best participant ARI
wiki-wiki	0.7096	2	0.9625
bts-rnc	0.2415	3	0.3384
active-dict	0.2144	3	0.2477

6. Conclusions

This is the description of our participation in the RUSSE'18 Russian Word Sense Induction shared task. We intended to create a very naive WSI system making use of pre-trained word embedding models and standard clustering algorithms. This enterprise was successful for the **wiki-wiki** dataset, but not so much for the **bts-rnc** and **active-dict** datasets: most probably, because they mostly consist of polysemous words with highly inter-related senses.

We showed that word embedding models trained on well-balanced and clean corpora (like the Russian National Corpus) can be superior in the extrinsic WSI task to those trained on large but noisy and unbalanced web or news corpora. This goes in line with the previous research which proved this for various intrinsic evaluation tasks.

The system we implemented is described in detail in this paper, and its Python source code is available online[6]. We hope that it will be of some use to other Russian NLP practitioners. Finally, we express our gratitude to the RUSSE'18 organizers for the chance to participate in an exciting shared task.

References

1. *Bar-Hillel Y.* (1964). Language and information; selected essays on their theory and application. Addison-Wesley.
2. *Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T.* (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5.
3. *Di Marco A., Navigli R.* (2013). Clustering and diversifying web search results with graph-based word sense induction. Computational Linguistics, 39(3)

[6] https://github.com/akutuzov/russian_wsi

4. *Droganova K. and Zeman D.* (2016). Conversion of SynTagRus (the Russian dependency treebank) to Universal Dependencies. Technical report, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University
5. *Jones, Karen Sparck* (1964). Synonymy and semantic classification. Edinburgh University Press.
6. *Kilgarriff, A.* (1997). I don't believe in word senses. Computers and the Humanities, 31(2).
7. *Kutuzov A., Kopotev M., Sviridenko T., Ivanova L.* (2016). Clustering Comparable Corpora of Russian and Ukrainian Academic Texts: Word Embeddings and Semantic Fingerprints, In Proceedings of the Ninth Workshop on Building and Using Comparable Corpora, held at LREC-2016. European Language Resources Association.
8. *Kutuzov A., Kuzmenko E.* (2017), WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, vol 661. Springer, Cham
9. *Kutuzov A., Kuzmenko E., Pivovarova L.* (2017). Clustering of Russian Adjective-Noun Constructions using Word Embeddings, In Lidia Pivovarova; Jakub Piskorski & Tomaž Erjavec (ed.), Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing. Association for Computational Linguistics. ISBN 978-1-945626-45-6.
10. *Kutuzov A., Kuzmenko, E.* (2016). Neural Embedding Language Models in Semantic Clustering of Web Search Results, In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016). ELRA.
11. *Kutuzov, A., & Kunilovskaya, M.* (2017). Size vs. Structure in Training Corpora for Word Embedding Models: Araneum Russicum Maximum and Russian National Corpus. In International Conference on Analysis of Images, Social Networks and Texts (AIST-2017). Springer, Cham.
12. *Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J.* (2013), Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems.
13. *Miller G.* (1995). Wordnet: a lexical database for English. Communications of the ACM, 38(11):39–41
14. *Moro, A., Navigli, R.* (2015). SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Association for Computational Linguistics.
15. *Navigli R., Ponzetto S.* (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence, 193:217–250.
16. *Panchenko A., Lopukhina A., Ustalov D., Lopukhin K., Leontyev A., Arefyev N., Loukachevitch N.* (2018): RUSSE'2018: A Shared Task on Word Sense Induction and Disambiguation for the Russian Language. In Proceedings of the 24rd International Conference on Computational Linguistics and Intellectual Technologies (Dialogue'2018).

17. *Panchenko, A., et al.* (2017). Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.
18. *Plungian V. A.* (2005), Why we make Russian National Corpus? [Зачем мы делаем Национальный корпус русского языка?], Otechestvennye Zapiski, 2.
19. *Schutze H.* (1998). Automatic word sense discrimination. Computational Linguistics Special-Issue-on-Word Sense Disambiguation, 24(1).
20. *Segalovich I.* (2003), A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine, MLMTA, pp. 273–280.
21. *Straka M., Straková J.* (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, August 2017.
22. *Alagić, D., Šnajder, J., Padó, S.* (2018). Leveraging Lexical Substitutes for Unsupervised Word Sense Induction. In Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18).
23. *Corrêa Jr, E. A., Amancio, D. R.* (2018). Word sense induction using word embeddings and community detection in complex networks. arXiv preprint arXiv:1803.08476.
24. *Navigli, R., Vannella, D.* (2013). Semeval-2013 task 11: Word sense induction and disambiguation within an end-user application. In Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)
25. *Bartunov, S., Kondrashkin, D., Osokin, A., Vetrov, D.* (2016). Breaking sticks and ambiguities with adaptive skip-gram. In Artificial Intelligence and Statistics (pp. 130–138).
26. *Benko, V., Zakharov, V.* (2016). Very large Russian corpora: New opportunities and new challenges. In Proceedings of the 22nd International Conference on Computational Linguistics and Intellectual Technologies (Dialogue'2016).

AUTOMATED TEXT READABILITY ASSESSMENT FOR RUSSIAN SECOND LANGUAGE LEARNERS[1]

Laposhina A. N. (antonina.laposhina@gmail.com),
Veselovskaya T. S. (tatianus2006@yahoo.com),
Lebedeva M. U. (m.u.lebedeva@gmail.com),
Kupreshchenko O. F. (ofkupr@gmail.com)

Pushkin State Russian Language Institute (Moscow, Russia)

This paper presents an outline of the readability assessment system construction for the purposes of the Russian language learning. The system is designed to help educators easily obtain the information about the difficulty level of reading materials. The estimation task is posed here as a regression problem on data set of 600 texts and a range of lexico-semantic and morphological features. The scale choice and annotated text collection issues are also discussed. Finally, we present the results of the experiment with learners of Russian as a foreign language to evaluate the quality of a predictive model.

Keywords: readability, text complexity, reading difficulty, graded readers

АВТОМАТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ СЛОЖНОСТИ РУССКОГО ТЕКСТА КАК ИНОСТРАННОГО

Лапошина А. Н. (antonina.laposhina@gmail.com),
Веселовская Т. С. (tatianus2006@yahoo.com),
Лебедева М. Ю. (m.u.lebedeva@gmail.com),
Купрещенко О. Ф. (ofkupr@gmail.com)

Государственный институт русского языка
им. А. С. Пушкина (Москва, Россия)

1. Introduction and related works

Today's information and text-rich world opens great opportunities for personalized learning, but at the same time, it sets the task of estimation and selection the suitable information. Suitable is understood as relevant to the educational purposes

[1] This research has been supported by the RFBR grant No.17-29-09156.

on the one hand and interesting and meaningful for this particular student on the other.

As R. Reynolds notes, tools for automatic identification of complexity of a given text would help to avoid one of the most time-consuming steps of text selection, allowing teachers to focus on pedagogical aspects of the process. Furthermore, these tools would also make it possible for learners to find appropriate texts by themselves [Reynolds, 2016].

In general, automated text difficulty assessment is the task of labeling a text with a certain difficulty level, such as grades, the age of the student, CEFR[2] levels, or some other abstract scale. The need of estimating texts by difficulty is not new: it starts from the beginning of the 20th century in a context of school education with quite simple formulas based on words and sentences length.

Nowadays both methods and possible application areas of such systems have widely expanded. Originated in the field of school education, researches on estimation of text complexity and search of appropriate ways of its simplification can play a significant role in specific applications where the accessibility of information is extremely important: for instance, readability assessing of government documentation for the general public[3], applications helping readers with dyslexia [Rello et al., 2012] or with intellectual disabilities [Feng et al., 2009], other groups of poor readers. Finally, the issue of finding educational texts with appropriate difficulty level for the second language learners is our particular interest. In modern NLP researches readability assessment posed as a data-driven machine learning task, is using a variety of text features from habitual word length to complex syntactic [Schwarm and Ostendorf, 2005] and discourse features [Pitler and Nenkova, 2008], features from statistical language models [Collins-Thompson and Callan, 2004], etc.

The task of text complexity estimation for the second language learners has some peculiarities. Thus, Heiman indicates the greater role of grammatical features in the second language readability research compared to native language one [Heilman et al., 2008]. The differences in the vocabulary level are also worth noticing. In our previous research [Laposhina, 2017] we have found out that the lexical group of features demonstrates one of the best correlation scores with text complexity in Russian. Perhaps, this is due to the difference in vocabulary acquisition of native and foreign languages. Walker et al. notes the disparity of reading in the native and the second or foreign language: when we first learnt to read in our first language, we already knew at least 5,000 words orally [Cunningham 2005], whereas we are usually plunged into reading a second language at an early stage, when we know very little of the language. L2 readers are constantly confronted with vocabulary they do not know [Walker, 2013]. Moreover, the differences in readability assessing for a second language include sufficiently clear and rigorous scale levels, knowledge and skill requirements for each level, word lists, and vocabulary.

There are a few readability researches for the Russian as a Foreign Language. R. Reynolds builds a six-level Random Forest classifier with a range of lexical,

[2] Common European Framework of Reference for Languages.

[3] https://plainlanguage.gov

morphological, syntactic, and discourse features and obtains F-score of 0.671. Better results were shown in binary classification task with two adjacent reading levels (e.g. A1-A2): F-score here is about 0.8–0.9. The author also provides information about feature`s information gain. [Karpov et al. 2014] use Classification Tree, SVM, and Logistic Regression models for binary classification of 4 CEFR levels (A1-C2, A2-C2, and B1-C2). The design of the given classification task seems not to fit the author's objective 'to retrieve appropriate material for their (students) language level' [Karpov et al., 2014], as the classification of adjacent reading levels is absent. A predictive model was trained on the base of 219 texts and 25 features including sentence and word length, the percentage of words from vocabulary lists and the number of several POS. The most predictive one were word lists. The authors also examine the sentence-level readability classification on 'B1 level and lower' and 'higher than B1' using transformed Dale-Chall model. [Sharoff et al. 2008] use Principal Component Analysis (PCA) in the aim to find the range of features that make a text difficult to read across a variety of languages without requiring complex resources, such as parsers. In order to realize that, they use word and sentence length, Flesch Readability Formula, average number of some specific word forms and coverage by frequency lists. The two main components from PCA can be interpreted as grammatical and lexical dimensions of difficulty. Authors also present the results of the experiment on using this system in actual language teaching.

2. Readability Assessment

As noticed by [Kevyn Collins-Thompson 2014], a machine-learning approach to readability prediction consists of three basic steps:

- First, a gold-standard training corpus of individual texts is constructed.
- Second, a set of features is defined that are to be computed from a text.
- Third, a machine learning model learns how to predict the gold standard label for a text from the text's extracted feature values.

Our work has been done in the established tradition. In section 1, the scale choice and training data set construction is discussed. Section 2 is devoted to feature extraction and selection; section 3 represents machine-learning algorithms training; and finally section 4 presents an evaluation experiment with a real educational life.

2.1. Scale choice and corpus constructing

For text complexity research a scale selection is being required: this will determine the way of corpus annotation and the type of machine learning task. Discussing traditional readability formulas, the text is considered to be suitable based on the reader's age or grade, but this differentiation does not reflect information about real reader's competence. This situation is clearly illustrated by the authors of the project on the personalization of the readability metrics Lexile[4]: in their video presentation

[4] https://lexile.com

they show a family who has come to a store to buy kid's sneakers; searching for a suitable pair, parents do not use child's individual shoe size, but focus on his age. The authors of this project offer an abstract numerical index that consists of text metrics and the vocabulary of a particular student as a scale.

An abstract scale is also widely used among readability studies: from 0 to 100 [Orphee De Clercq, 2017], 1 to 5 [Pitler and Nenkova, 2008], binary—easy/difficult or suitable/not suitable for this level, triple—simple/average/difficult [Selegey et al., 2015]. Quite easy and effective way to get annotated training data may be using parallel collections of texts: e.g. Simplified VS Normal Wikipedia, [Sharoff et al, 2008], Children VS Adult version of Encyclopedia Britannica [Schwarm and Ostendorf, 2005]. Regarding the multi-level scale, it can be graded reader collections such as Weekly Reader, an educational newspaper with texts targeted at different grade levels [Weekly Reader, 2004].

As for a second language readability studies, the most common decision here is to use a standard grading scale for foreign language proficiency which is already developed for assessment and certification of foreign students ([Reynolds, 2016]; [Karpov, 2014]; [Schwarm and Ostendorf, 2005]). For European languages, this is the CEFR[5] level system that is measured with a six-level scale from A1 (Beginner) up to C2 (Proficiency). This system of levels has several advantages:

1. The availability of the specific regulatory documents that clarify the requirements for knowledge of vocabulary, grammar and syntax for each level.
2. Independence from such subjective categories as grade / age / number of years of study. These levels have a specific amount of language material that a person who claims to have a certificate of appropriate level should know.
3. The textbooks contain information on what levels they are intended.
4. There is a correlation with the real-life situations (for example, it is necessary to have a certain level to enter Russian universities, get a job in Russia, get Russian citizenship, get permission to teach Russian, etc.). Thus, the level of text complexity becomes a less abstract category.

level	A1	A2	B1	B2	C1	C2	C2+
number of text	108	120	106	97	39	75	48

Table 1. *Corpus content distribution*

There are 6 basic CEFR levels, translated into numerical form (A1 = 0, A2 = 1 etc) at the core of our scale. So the complexity of the text is presented as an increasing value, that reflects the concept of process of language acquisition more naturally, than 6 closed classes. Our corpus contains about 600 texts from the CIE resource[6] and several textbooks. Authors of these books provided the information about the target level. The content distribution is shown in **Table 1**.

[5] https://en.wikipedia.org/wiki/Common_European_Framework_of_Reference_for_Languages

[6] http://texts.cie.ru

For the C2 (the level of an educated native speaker) texts from news portals and articles from popular magazines on various subjects were used. However, it is obvious that the reading difficulty of texts marked as native speaker level can also differ greatly. Therefore, we added to our scale the C2+ level, which will include texts supposedly perceived as difficult by Russian native speakers: texts of laws, articles from the popular science magazine N+1[7], noted by the editors as complicated (they defined complexity as an amount of the scientific background in this field which is needed to understand the article).

We have faced several issues while collecting corpus: a) information about level could be absent in the textbook; b) this information may not contain a clear indication of the CEFR levels (B1-B2, «advanced», «for the second semester»); c) the difficulty level reflects the author's subjective evaluation. Therefore, in the future we are planning to perform expert or crowdsourcing annotation of our text collection to fix these limitations and to get more objective information about complexity of given texts using average score from several annotators.

2.2. Feature extraction and selection

We select the features to extract taking into account the following principles: 1) the features should reflect the information provided in the regulatory documents. 2) following [Sharoff et al. 2008], we believe that the features should be quite simple and reproducible if we are talking about the real usage of this system in language learning.

First, we extract some basic text metrics such as average and median word length, sentence length, average number of syllables per word, percentage of 'long' words (more than 4 syllables), average number of punctuation marks per sentence. This group of features is easy to get but it is still capable to show high correlation with a difficulty level in an obvious way: the longer the text and the words in it, the more likely it is difficult to read.

[François and Miltsakaki 2012] in their readability study have found that the best prediction performance was obtained using both classic (readability formulas) and non-classic features. Considering this, we have applied as a feature 5 commonly used readability formulas, which are using following parametres:

1. Flesch–Kincaid: (words / sentences) + (syllables / sentences);
2. Coleman Liau index: (characters / words) + (sentences / words);
3. Automated Readability Index: (characters / words) + (words / sentences);
4. Dale-Chall formula: ('difficult' words that are out of Dale's 3000 simple words list / all words) + (words / sentences);
5. Simple Measure of Gobbledygook: (words more than 4 syllables / sentences).

More information about readability formulas adaptation for Russian is available in [Begtin, 2015].

Following previous researches (e.g. [Pitler and Nenkova, 2008]; [Zeng et al., 2008]; [Laposhina, 2017]), we paid attention to the group of lexical features: there

[7] https://nplus1.ru

are subsets of features based on coverage by vocabulary lists for each level ("lexical minimums"), frequency lists by Lyashevskaya and Sharoff[8] and Brown [N. Brown, 1996], and number of words from some specific word lists: abstract words, emotional words, verbs of motion, modal constructions, Dale's list of 3000 "simple words"[9], lists of 1000 and 2000 basic words from the Basic English Project[10]. As for the last ones, we realize the roughness of the English word lists` translation, but even approximate information on their correlation to the text complexity in Russian can motivate our further study in this field.

The next feature subset provides data about grammatical information. The percentage of POS or grammatical forms is counted here for a sentence and for a whole text, e.g. 'percent of nouns in a sentence', 'percent of nominative case in a text'.

To estimate the impact of these features in Russian second language readability assessment, the Pearson and Spearman correlation coefficients and p-value were calculated[11]. Top-30 features contains all groups of features but in different proportions.

Feature	Pearson coefficient	p-value	Spearman coefficient	p-value
A2 word list coverage of a text	-0.85	1.3e-171	-0.87	5.6-e186
Formula SMOG	0.75	2.6e-110	0.74	6e-108
Mean sentence length	0.72	3.6e-100	0.71	1.1e-96
10000 frequency word list coverage of a text	-0.69	2.2e-86	0.70	1.3e-90
Dale 3000 word list coverage of a text	-0.68	1.6e-84	0.70	1.3e-92
Abstract words list coverage of a text	0.58	3.9e-57	0.60	2.3e-63
Percentage of neuter words per text	0.55	1.3e-49	0.60	5e-68
Median number of punctuation per sentence	0.55	1.4e-49	0.55	7.3-50
Percentage of words in genitive case per text	0.50	2.8e-39	0.60	6.1e-60

Table 2. *Examples of correlation coefficient for different groups of features*

The highest correlation was shown by the lexical minimums coverage—this fact not only confirms the connection between the lexical minimums and text difficulty, but also characterizes the corpus content, which consists mostly of the textbooks, designed, in turn, according to the lexical minimums,—this has led to a vicious circle. All five readability formulas, sentence and word length information have also shown top results. The presence of features from specific word lists as Dale Word List and Basic English translated versions encourage us to continue research in this area and to develop similar lists for Russian.

The top morphological features are presented by the percentage of neuter nouns, words in nominative and genitive cases, and participles. Most of the grammatical

[8] http://dict.ruslang.ru/freq.php

[9] https://en.wikipedia.org/wiki/Dale–Chall_readability_formula

[10] http://ogden.basic-english.org/

[11] https://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.pearsonr.html

features have positive correlation (e.g. the high proportion participles can indicate passive forms and specific Russian participle constructions which cause difficulties in understanding among foreigners; a number of neuter nouns may be connected with a number of special terms and abstract concepts). In contrast, there is a negative correlation between the percentage of words in nominative case and the difficulty level, as the less often the nominative case occurs in the sentence, the larger a proportion of the oblique cases is in it, which is also difficult. Mean sentence length, number of prepositions and conjunctions may indicate a syntactical aspect of difficulty.

A number of linear correlations between these text features was detected, e.g. connection between different readability formulas or lexical minimums, frequency lists. We will keep it in mind while model fitting.

2.3. Regression model

The aim of this part of work was to predict the correct assessment of a text from the continuous-valued scale from 0(A1) to 6(C2+). In order to do this, we have experimented with two linear regression algorithms: ordinary least squares Linear Regression and Ridge Regression (linear least squares with l2 regularization) by scikit-learn[12]. The mathematical objective of this techniques is to minimize mean squared error.

	Linear Regression	Ridge regression
all 149 features		
explained variance	0.73	0.83
mean squared error	0.67	0.49
44 best correlation features		
explained variance	0.82	0.84
mean squared error	0.49	0.46

Table 3. *Model Evaluation*

The models were built with:

a) all 149 features

b) 44 features with correlation by Pearson > 0.3.

To evaluate the results we have used a standard metrics as explained variance score and mean squared error. The best result was achieved by Ridge regression based on 44 best correlation features. We assumed that Ridge Regression better results may be explained by its resistance to multicollinearity of features. Twenty-fold cross-validation test showed accuracy 0.82 (±0.05) for Ridge Regression and 0.80 (±0.07) for Linear Regression.

To visualize the output of an algorithm a confusion matrix was constructed. Rows represent here an actual level by the corpus data while columns represent predicted levels.

[12] http://scikit-learn.org

Standart \ Prediction	A1	A2	B1	B2	C1	C2	C2+
A1	21	3	0	0	0	0	0
A2	9	15	4	1	0	0	0
B1	1	7	12	0	0	0	0
B2	0	5	9	17	5	0	0
C1	0	0	0	3	4	0	0
C2	0	0	0	7	14	3	0
C2+	0	0	0	0	2	5	3

Table 4. *Confusion matrix*

Table 4 shows, that mispredictions more than 1 level are only 10% of a test set, that is quite encouraging. It's also interesting to note 'the direction' of errors: algorithm more often underestimates the difficulty (47 VS 12), especially at high levels. One of the reasons of such phenomenon may be connected with the peculiarity of the corpus content: texts in B2 and C1 textbooks are aimed at confident users of Russian and provide information on complex grammatical constructions and various functional styles of the Russian language, so they can be more difficult, than the usual news articles that we collected for the C2 level.

Text	*Predicted level*
Tale story 'Masha and bears'	2.2 (B1)
Article from travel blog (1 000 words)	2.9 (B1)
A. Chekhov. Basic Education (a novel)	3.1 (B2)
A. Pushkin. The Captain's Daughter (3 000 words)	4.2 (C1)
The contract for renting an apartment	4.5 (C1)
V. Nabokov. Lolita (3 000 words)	5.9 (C2)

Table 5. *Example of predictions*

The examples of system working with authentic Russian texts are shown below. These results correspond both to the intuition of the expert teachers and to the requirements of the state standards for Russian learners, where reading authentic texts with minimal adaptation are appropriate for readers beginning with B1 level and above. More details on the evaluation experiment will be described in the next section.

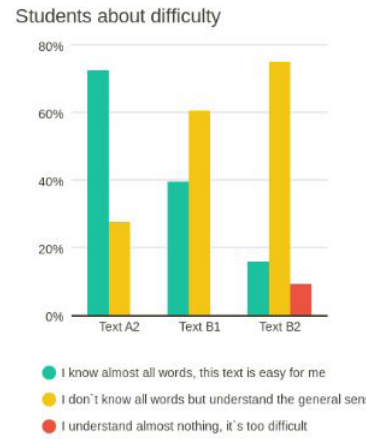
2.4. Evaluation

To test the accuracy of our approach to automatic text complexity measurement and to estimate its applicability in real educational life, we have proceeded an experiment with 78 international students at B1 level of Russian language proficiency. It took place at the Pushkin State Russian Language Institute in February 2018. Three authentic texts on the similar topic with minimal adaptation were prepared; our system evaluated them as A2, B1 and B2 respectively. The students were asked to read

each text without dictionary, to mark unknown vocabulary, to do post-reading quiz and to note how difficult to understand these texts were.

	Text A2	Text B1	Text B2
Level assessement by the algorithm	1.63	2.69	3.02
Level assessement by teachers	1.17	1.7	2.35
Median reading time (minutes)	4	5	6
Number of words out of B1 word list	11 (6%)	21 (10%)	31 (15%)
Mean number of words marked by students as unknown	0.9	3.11	5.9
Students anwered correct at least 2 of 4 post text quiz	96%	88%	76%
Students answered correct all 4 post text quiz	42%	20%	15%

Table 6. *Survey results*



The core insights from this study are shown below. The scale of text difficulty is readily seen here: the more difficult by our algorithm text is, the more words and syntactic constructions are marked by students as unknown and the less percentage of correct quiz answers are given. During personal interview students also easily ordered given texts by difficulty level highlighting that text 3 is the most difficult. Besides, they avoided to pick the option 'I understand almost nothing, this text was too difficult' in the questionnaire: this can be caused both by psychological factors, when intermediate-level students are not comfortable to admit such an overgeneralized option and by weakness of our program due to it's tendency to overestimate the real level of the text difficulty. We will take it into account while our further research.

3. Conclusion and further work

In this article we presented a supervised approach for text complexity assessment for Russian as a Second Language using linear regression. The best result was performed by Ridge Regression algorithm, trained on the 44 best correlation features set. As our further work we can point out such directions as:

1. Corpus expansion, adding the segment with authentic texts, mainly annotated by different experts;
2. Searching for new lexico-semantic features (polysemantic words, idioms and collocations, archaisms and historicisms, conversational vocabulary, genre-specific words seem particularly promising).

References

1. *Begtin, I. V.* (2014), What is "Clear Russian" in terms of technology. Let's take a look at the metrics for the readability of texts: the blog of the company "Information Culture" [Chto takoe "Ponjatnyj russkij jazyk" s tochki zrenija tehnologij. Zagljanem v metriki udobochitaemosti tekstov: blog kompanii "Informacionnaja kultura"], available at: http://habrahabr.ru/company/infoculture/blog/238875/
2. *Brown, N.* (1996), Russian Learners' Dictionary: 10,000 Russian Words in Frequency Order, Routledge, 1996.
3. *Collins-Thompson, K.* (2014), Computational assessment of text readability: a survey of current and future research. In: François, Thomas and Delphine Bernhard (eds.), Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics 165:2, pp. 97–135.
4. *Collins-Thompson, K., & Callan, J.* (2004), A language modeling approach to predicting reading difficulty. In Proceedings of HLT-NAACL 2004, pp. 193–200.
5. *Feng, L., Elhadad, N., & Huenerfauth, M.* (2009), Cognitively motivated features for readability assessment. In The 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009), pp. 229–237.
6. *François, T., Miltsakaki, E.* (2012). Do NLP and machine learning improve traditional readability formulas? Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations, Association for Computational Linguistics, 2012, 49–57.
7. *Heilman, M. J., Collins, K., Callan, J., & Thompson, M. E.* (2007), Combining lexical and grammatical features to improve readability measures for first and second language texts. Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference, Rochester, New York, USA, pp. 460–467.
8. *Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M.* (2007), Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. In Proceedings of HLT-NAACL'07, pp. 460–467.
9. *Karpov N., Baranova J., Vitugin F.* (2014), Single-sentence readability prediction in Russian. In Proceedings of Analysis of Images, Social Networks, and Texts conference (AIST), pp. 91–100.
10. *Laposhina, A.,* (2017), Relevant features selection for the automatic text complexity measurement for Russian as a foreign language. [Analiz relevantnyh priznakov dlya avtomaticheskogo opredeleniya slozhnosti russkogo teksta kak inostrannogo], Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2017), Issue 17, p.1–7.
11. *Pitler, E. & Nenkova, A.* (2008), Revisiting readability: a unified framework for predicting text quality. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 186–195.
12. *Rello, L., Saggion, H., Baeza-Yates, R., Graells, E.* (2012), Graphical schemes may improve readability but not understandability for people with dyslexia. Proceedings of NAACL-HLT 2012.

13. *Reynolds R.* (2016), Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories, San Diego, CA: 16 June 2016. In: Proceedings of the 11th Workshop on the Innovative Use of NLP for Building Educational Applications, pp. 289–300.
14. *Schwarm, S. E. & Ostendorf, M.* (2005), Reading level assessment using support vector machines and statistical language models. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL '05). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 523–530.
15. *Sharoff S., Kurella S., Hartley A.* (2008), Seeking needles in the web's haystack: Finding texts suitable for language learners. In Proceedings of the 8th Teaching and Language Corpora Conference, (TaLC-8), Lisbon, Portugal.
16. *Walker A., White G.* (2013), Technology Enhanced Language Learning: connecting theory and practice, Oxford University Press.
17. *William H. DuBay* (2006), The Classic Readability Studies. Impact Information, Costa Mesa, California.

LEXICAL VARIATION: WORD KNOWLEDGE AND POLYSEMY IN RUSSIAN EVERYDAY LIFE LEXICON[1]

Levin I. (levinivanse@gmail.com),
Andriyanets V. (blindedbysunshine@gmail.com)
National Research University "Higher School of Economics"

Iomdin B. (iomdin@ruslang.ru)
V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences; National Research University "Higher School of Economics"

Ambartsumian A. (anna.ambr@yandex.ru)
Russian State University for the Humanities

Many words that according to the dictionaries have just one meaning are in fact understood in different ways by different speakers. In this article we deal with Russian nouns denoting everyday life objects which are subject to much variation by age, gender, and region and are poorly described by the existing dictionaries. We report the results of a multilevel survey, propose some possible metrics of word knowledge and show to what extent the words we studied are known among a certain population. We also claim that different speakers possess different sets of meanings for each word, propose ways to discover the distribution patterns for these sets and introduce the notion of disperse polysemy. We believe that our findings may be useful in lexicography (providing detailed information on current word usage in different social groups), lexical semantics (researching meaning shifts and patterns of its distribution among speakers), and language testing (more precise detection of the vocabulary sizes both in native speakers and in language learners).

Key words: semantics, polysemy, lexicography, lexical variation, word knowledge

0. Introduction

Linguists and lexicographers often deal with polysemy. In natural language processing, in particular, a lot of research is aimed at word sense disambiguation, normally context-dependent [Ide and Veronis 1998, Navigli 2012, Chandra and Dwivedi 2014, Iomdin 2014, Iacobacci et al. 2016]. However, many words that according to the

[1] The research of Boris Iomdin was supported by RSF (project No. 16-18-02054: Semantic, statistic and psycholinguistic analysis of lexical polysemy as a component of Russian linguistic worldview).

dictionaries have just one meaning are in fact understood in different ways by different speakers. We are currently researching this issue as part of our work on the Thesaurus of Russian Everyday Life Lexicon [Iomdin 2011] and within the project "Semantic, statistic and psycholinguistic analysis of lexical polysemy as a component of Russian linguistic worldview" funded by RSF.

Here we deal with nouns denoting everyday life objects which are subject to much variation by age, gender, and region [Iomdin 2014] and are very poorly described by the existing dictionaries, so we have to obtain the necessary data from sociolinguistic surveys. Corpora are less useful here, because in the texts that they incorporate artifacts are rarely described in detail sufficient to distinguish between similar objects and to provide accurate and distinctive definitions. Since we started working on the thesaurus we have conducted many surveys to this aim. In some of them, we asked the respondents whether they knew certain Russian words, and the analysis of the results clearly shows that the answer to this question cannot be binary.

Various experiments dedicated to detecting the vocabulary size of native speakers were conducted based on the idea that one can obtain a specific number of words known by the respondent because each word can be assumed either known or unknown. In one of the earliest experiments of this type, [Hartman 1946] related the idea of *knowing* the word with the ability to give a definition to it. One word was selected from every fortieth page of Merriam Webster's New International Dictionary, and so a list of 50 words was created. On average, the students were able to define 26.9 out of the 50 words, a proportion that gave the impression that they had a vocabulary of 215,000 English words. [Goulden 1990] used the same dictionary to choose the words for their experiment, but the final list was reduced by excluding proper nouns, derived nouns and compounds. He presented lists of 50 words each to 20 university graduates who had to indicate whether they knew the word without proving it. The result was 17,200 known English words, lower than in previous experiments. In [Milton and Treffers-Daller 2013] some parts of the two previous studies were united: they took the reduced list of words from [Goulden 1990] and asked first-year university students to provide either a definition or a synonym for each word they knew. In this experiment, the resulting figure was still lower: 9,800 word families (morphologically related groups of lemmas) known to an average respondent. These three experiments show that the more complicated the selection of lemmas for the final list and the structure of the experiment are, the less is the number of known words that is received as a result.

Along with the selection of words, the fact that a word can have multiple senses or meanings must be taken into account. [Rodd, Gaskell, Marslen-Wilson 2002] studied the response time in a lexical decision task and its correlation to the numbers of senses or meanings. The results of three experiments represented "an important challenge to accepted views of how semantic ambiguity affects recognition of isolated words. Ambiguity between multiple meanings produces a disadvantage, while multiple senses produce faster responses". [Brysbaert and Stevens 2016] in their work dedicated to the same issue (how many words a native speaker knows) note the importance of qualitative estimation of the result: "our assessment says nearly nothing about how well the participants know the various words".

The present work is based on the hypothesis that word knowledge may have a more complicated structure which includes various levels. In order to test this hypothesis, we conducted a multilevel survey, for which we selected seven Russian nouns denoting everyday objects. In **Section 1** we deal with the history and semantic development of these nouns. In **Section 2** we describe the design of the experiment. In **Section 3** we propose possible approaches to defining word knowledge and provide corresponding data from our experiment. In **Section 4** we discuss the question whether we deal with polysemy in the cases we studied.

1. Material

Having analyzed Russian text corpora (mainly Russian National Corpus and RuTenTen11) as well as the Google Books collection and various online resources, we selected several less frequent Russian words that apparently are understood differently by different speakers.

Сланцы [*slancy*] 'flip-flops, jandals'. The etymology of this word can be traced to a proper name. The factors contributing to its meaning development were probably (1) the plural form of the word, characteristic for all kinds of shoes, (2) its similarity to an older word *šljopancy* 'sliders, jandals', and (3) the novel nature of this kind of shoes and lack of a conventional name for it (another frequent colloquial term for them are *v'etnamki*, lit. 'Vietnamese', absent in dictionaries but appearing in published texts since the 1970s). The meaning shift is an example of metonymy (a label on the object → the object itself).

Барсетка [*barsetka*] 'man bag, man purse, murse'. This word is in all probability borrowed from Italian, where *borsetta* (and *borsetto*) is a diminutive form of *borsa* 'bag'. The meaning, however, differs from the Italian word *borsetta*, which means 'women purse', whereas the meaning of the Russian word is similar to that of *borsello* 'a small bag for men with the function similar to that of the female handbag, often with a strap that allows one to hang it on the shoulder'. Apparently, it was not the name of this very object, but rather the label applied to various kinds of leather handbags that was used as the basis for the Russian word.

Креманка [*kremanka*] 'ice-cream bowl, dessert bowl'. This word is generally considered to be a derivative of *krem* 'cream, hard sauce', which has a common meaning component ('dessert'). However, the word *kremanka* is now normally used for a bowl for ice cream and other desserts but not for hard sauce served separately. Moreover, the suffix *-ank(a)* is common for animated nouns rather than inanimate nouns derived from names of objects or substances. We believe that *kremanka* is a derivation not from *krem*, but from *kreman*, a now obsolete word borrowed from the French *crémant* 'sparkling wine'. Examples of *kremanka* used in this sense can be found in texts published in late 19th century and early 20th century. The term was used for champagne coupes, which at some point started to be used as ice-cream bowls. Here we deal with a metaphorical shift: an object got its name

from another object with a different purpose, but of the same form; its phonetic resemblance to the word *krem* associated with desserts contributed, too.

Тренч [*trenč*] 'trench, trench coat'. This word apparently was borrowed from English twice. English dictionaries list two senses for *trench coat*: (1) usually double-breasted raincoat with deep pockets, wide belt, and often straps on the shoulders, and (2) a waterproof overcoat with a removable lining designed for wear in trenches". The word *trench coat* was borrowed as a whole, occurring in Russian texts in the 1930s (first *трэнчкот*, then *тренчкот*). Then, according to RNC and Google Books, it was rarely used until a rebirth at the beginning of the 2000s, normally as just *тренч*. It is now associated with youth fashion rather than military style.

Манто [*manto*] 'fur opera cloak'. This word was borrowed from French *manteau* at the beginning of the 19th century (at first it was masculine, then neuter). It used to mean a coat in general, particularly a light one. Then its meaning narrowed down to women fur opera cloaks.

Душегрейка [*dušegrejka*] 'a warm women jacket', literally 'soul warmer'. This word used to describe a traditional women outer garment, generally sleeveless and warm. Another word of similar structure, *telogrejka*, literally 'body warmer', according to several dictionaries, was used as a synonym to *dušegrejka*. Later, however, the meanings of both words started to differ, and now *dušegrejka* often describes a fashionable women garment, whereas *telogrejka* is used to describe a warm cotton quilted jacket used in the Soviet army and labor camps. This divergence of meanings may be connected with different associations of *duša* 'soul' as something fragile vs. *telo* 'body' as something earth-bound.

Трюмо [*trjumo*] 'console mirror, three-leaved mirror'. This word was borrowed from the French *trumeau*. The Russian dictionaries list two senses: (1) 'console mirror, standalone mirror', (2) 'trumeau, pillar' (in architecture). However, the word frequently refers to a three-leaved mirror. Here, again, at least two factors contributed to this meaning development: (1) the advancement of three-leaved mirrors and the lack of a one-word nomination for such an object, (2) the phonetic similarity to the word *tri* 'three'. The latter factor also influenced the development of another word, *trel'jaž*, which is now a close synonym to *trjumo*. *Trel'jaž*, too, was borrowed from the French *treillage*. The French word means 'trellis, latticework', and one of the senses of the Russian word is close to it. However, no later than in the 1930s the word acquired a new, now much more frequent sense 'console mirror, three-leaved mirror'. The factors contributing to this meaning development are probably exactly the same as in the case of *trjumo*, even though the reason of its closeness to the root *tri* is entirely different. For a given speaker of contemporary Russian, the two types of mirrors can be expressed by these two words (*trjumo* and *trel'jaž*), the former meaning 'a console mirror' and the latter 'a three-leaved mirror', or vice versa: a distinction not quite described by the dictionaries.

We can see here different kinds of meaning shifts, various situations of borrowings, paronymic attraction and influence of official stock lists, inventories and industrial naming practices, resulting in rather complex and varying sets of meanings, which we decided to investigate further through a sociolinguistic survey.

2. Experiment design and participants

The questionnaire[2] was organized as follows. For each of the seven words, the participants were first asked to identify the only existing word among four possible options. The options for the word *barsetka*, e.g., were *barsetka*, *barfetka*, *baržetka*, and *barzetka*. Afterward, the participant had to choose which semantic field the word belonged to (they were presented with four options such as clothing, food, crockery, etc.), and then to choose the nearest hypernym from four given options. These three stages were used to detect how familiar the words in question are to the participants. One could proceed to the following stage only if they chose the correct option in the previous one.

Finally, in the fourth stage, the participants were presented with four pictures that represented different variants of the objects in question. Every picture was accompanied by a short description. Unlike the previous stages, this one was a multiple-choice task and had no presupposed correct answers.

1706 people participated in the study, including 1297 (76%) women, 404 (24%) men, and 5 people who did not submit the information on their gender. The median age of the participants was 32 years. We grouped the places of residence indicated by the participants into the following regions (see the map on **Figure 1**):

- Russia: Moscow (917 participants; 53,8%), Moscow Oblast (4,6%), Saint Petersburg (9,2%), the Center (11,1%), the East (9,5%), the South (1,4%), the North-West (1,8%).
- Ukraine (2,3%), Belarus (1,2%), and other countries (2,3%).

Fig. 1

[2] Available at https://goo.gl/forms/fYqG0aHHW2hOC1w73

3. Defining word knowledge

A person may know a word passively or actively, understand its meaning with certain precision and be familiar with a certain set of meanings if a word is polysemous. Here we propose some possible metrics of word knowledge that can be identified based on the data from the first three stages of our experiment and show to what extent a word is known among the participants.

These metrics are presented in Table 1. The second column represents the percentage of people who can identify a word as the only existing one among several similar strings of letters, i.e. the percentage of people who at least know that such a word exists in the language. The word *slancy* has the highest score while *kremanka* has the lowest one. The third column shows the percentage of people who know to what semantic field the word belongs, with *slancy* having again the highest score and *trenč* having the lowest score. Finally, the percentage of people who can give the nearest hypernym is presented in the fourth column. Again, *slancy* has the highest score and *trenč* and *kremanka* have the lowest scores.

While the general ranking of the seven words is more or less the same, the differences between different metrics vary to a remarkable degree. These differences are given in the last two columns of Table 1. N1–N2 shows the number of participants who correctly identified the word among similar nonce words but have no idea what it actually means. N2–N3 shows the number of participants who only know the meaning roughly.

Table 1. Different degrees of word knowledge

Word	Word identification among similar nonce words (N1)	Correct semantic field (N2)	Correct closest hypernym (N3)	N1–N2	N2–N3
slancy	99,4%	94,4%	93,7%	4,0%	0,7%
trjumo	97,4%	94,2%	84,6%	3,2%	9,6%
barsetka	95,2%	91,2%	83,0%	4,0%	8,2%
manto	91,5%	89,5%	74,4%	2,0%	15,1%
dušegrejka	87,1%	83,3%	72,6%	3,8%	10,7%
trenč	83,1%	74,6%	70,7%	8,5%	3,9%
kremanka	84,1%	78,4%	70,6%	5,7%	7,8%
mean	**91,1%**	**86,5%**	**78,5%**	**4,5%**	**8,0%**

A sociolinguistic dimension can make this picture even more complex. Below we show how our third knowledge metric (N3) depends on the social variables (gender, age, and region) using the words *kremanka* and *trjumo* as examples.

The decision tree for the knowledge of the word *kremanka* is presented in Figure 2. It divides the participants of the experiment into several groups based on the social variables. Figure 3 shows the analogous data for the word *trjumo*. The variation among the groups is not so striking. Still the difference is statistically significant.

This kind of data is missing from traditional dictionaries and can be useful for understanding that the prevalence of a particular word is not a simple quantitative variable but rather a complex entity that involves different social factors. As a whole, based on our findings, at least in the domain of everyday life lexicon, females and older people are more likely to know the words in question.

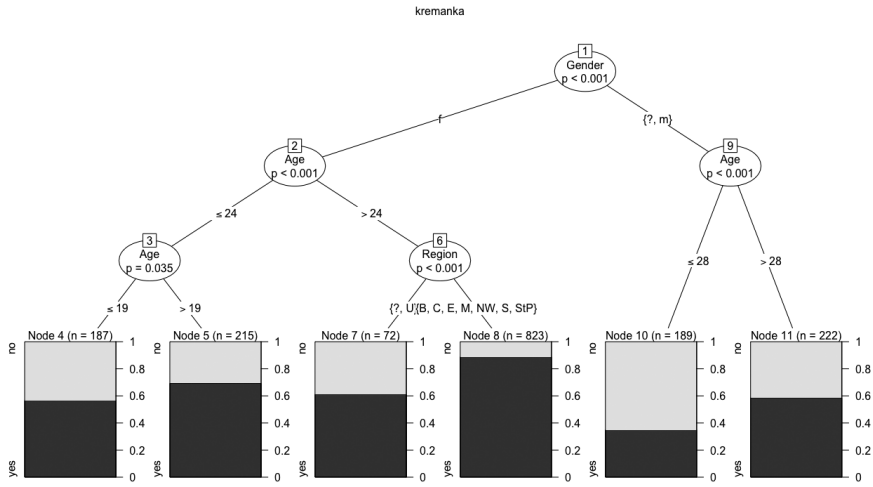


Fig. 2. Decision tree for the knowledge of the word *kremanka*

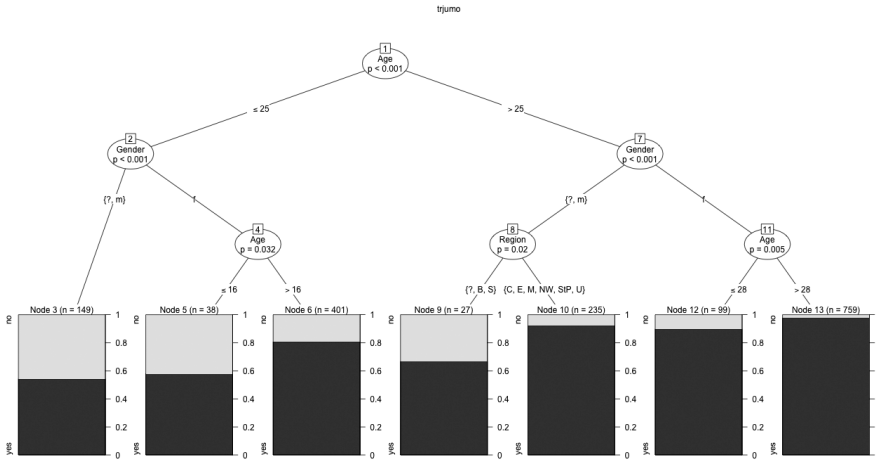


Fig. 3. Decision tree for the knowledge of the word *trjumo*

4. Is it polysemy?

One and the same word denoting an everyday life object can refer to objects different in shape, dimensions, or function. In each case, lexicographers have to decide whether to describe these differences as different dictionary senses. Here are the English translations of the sets of descriptions proposed in our experiment for the objects in question.

Slancy: (1) beach footwear with a strap between the toes; (2) beach footwear with a strap across the foot; (3) street footwear with a strap between the toes; (4) street footwear with a strap across the foot.

Barsetka: (1) man purse with a loop handle around the wrist; (2) small briefcase; (3) wallet with many pockets; (4) belt bag.

Kremanka: (1) small dessert bowl with a stem; (2) small dessert bowl without a stem; (3) small salad bowl with a stem; (4) small salad bowl without a stem.

Trenč: (1) city coat with pockets and a belt; (2) military style coat with a wide belt; (3) raincoat; (4) military coat without a belt.

Manto: (1) short elegant sleeveless fur coat; (2) long elegant fur coat with sleeves; (3) fur mantle; (3) light coat.

Dušegrejka: (1) ethnic Russian women jacket; (2) fur waistcoat; (3) cotton quilted jacket; (4) jacket with a fur collar.

Trjumo: (1) alone standing mirror; (2) dresser with a mirror; (3) dresser with three mirrors; (4) table with three mirrors.

Participants of our experiment were given these descriptions along with the pictures of these objects and were free to choose any set of them, including none or all of them. In most cases, participants chose only one option for each object, and these choices were significantly different. This may mean that most speakers have clearly defined mental images, rather than fuzzy concepts, behind these words, but these images differ across the pool of participants. This can hardly be considered true polysemy because the meanings are quite close to each other: in most cases, they have the same genus proximum, the same hypernyms and nearly the same synonyms. On the other hand, we could not provide a common definition for each meaning set in each group, because it would invariably be way too vague and broad. We would call this a case of *disperse polysemy*: a situation where several close but distinct definitions can be assigned to a word, which hardly ever coexist in a single speaker's mind, but rather in the speakers' population as a whole. Upon analyzing the distribution of these meanings, we can list them in the dictionary entry assigning labels with sociolinguistic information.

As an example of a possible analysis of how the four meanings of the word *trjumo* are organized, we provide a decision tree model that takes into account all social variables and the subsets of usages ascribed. This particular model (see **Figure 4**) shows with what probability people acknowledge that a dresser with three mirrors can be called *trjumo*. The lowest percentage (Node 8) corresponds to people older

than 26 years that acknowledge that an alone standing mirror can be called *trjumo*. It can be explained by the fact that this meaning is diachronically and semantically the most remote one from the one in question for this model. It also means that these two meanings are almost incompatible for the word *trjumo* within the lexicon of a single person (at least among the corresponding age group). The highest percentage (Node 5) is found among the people not older than 26 years who acknowledge also 'dresser with one mirror' and 'table with three mirrors', i.e. closely related meanings to the meaning in question, as possible meanings of *trjumo*.

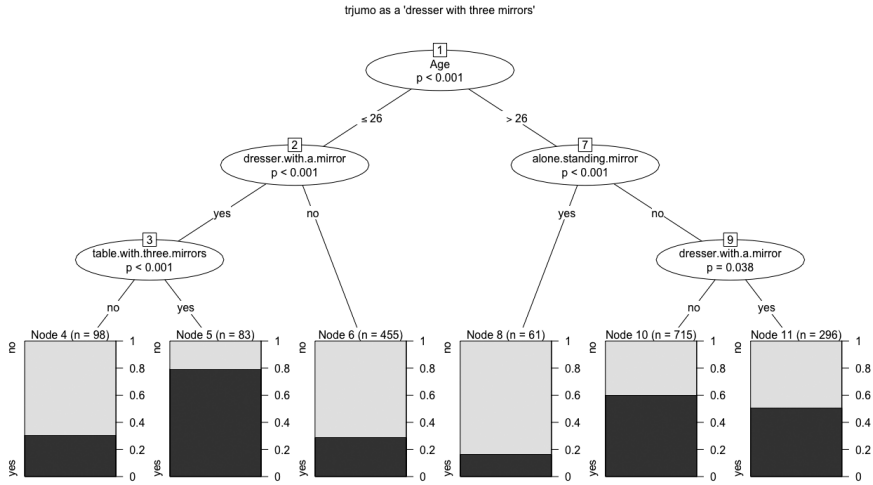


Fig. 4. Decision tree for *trjumo* as 'a dresser with three mirrors'

5. Conclusion

We believe that our findings may be useful in lexicography, lexical semantics, and language testing.

The existing dictionary entries are often insufficient and too narrow. For the purposes of our thesaurus, we intend to take into account the lexical variation and include different descriptions into the lexical entries, thus providing the dictionary users with more accurate and detailed information on current word usage in different social groups.

While we believe that the proposed notion of disperse polysemy is most characteristic for concrete nouns referring to artifacts, it can be further verified on various kinds of lexemes.

What we have shown in this paper are just several excerpts of the vast data that we collected. A further analysis can be conducted to see whether there is a correlation between participant profiles and the number of meanings they know; we could hypothesize that certain groups of respondents are better at handling polysemy than others.

The design of the multilevel survey that we created can be used for more precise testing of the vocabulary sizes both in native speakers and in language learners, if applied to mass lexical material, including much more frequent words.

References

1. *Brysbaert M., Stevens M., Mandera P. and Keuleers E.* (2016). How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. Front. Psychol. 7:1116.
2. *Chandra, G., & Dwivedi, S. K.* (2014). A literature survey on various approaches of word sense disambiguation. In Computational and Business Intelligence (ISCBI), 2014 2nd International Symposium on (pp. 106–109). IEEE.
3. *Goulden, R., Nation, I. S. P., and Read, J.* (1990). How large can a receptive vocabulary be? Appl. Linguist. 11, 341–363
4. *Hartmann, G. W.* (1946). Further evidence on the unexpected large size of recognition vocabularies among college students. J. Educ. Psychol. 37, 436–439
5. *Iacobacci, I., Pilehvar, M. T., & Navigli, R.* (2016). Embeddings for word sense disambiguation: An evaluation study. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Volume 1: Long Papers, pp. 897–907.
6. *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.
7. *Iomdin B. L.* (2014). Polysemous words in and out of the context. [Mnogoznachnyje slova v kontekste i vne konteksta]. Voprosy jazykoznanija [Issues in Linguistics]. Vol. 4. Moscow.
8. *Milton, J., and Treffers-Daller, J.* (2013). Vocabulary size revisited: the link between vocabulary size and academic achievement. Appl. Linguist. Rev. 4, 151–172
9. *Navigli, Roberto.* (2012). A quick tour of word sense disambiguation, induction and related approaches. In: International Conference on Current Trends in Theory and Practice of Computer Science. Springer Berlin Heidelberg.
10. *Rodd, Gaskell, Marslen-Wilson.* (2002). Making Sense of Semantic Ambiguity: Semantic competition in Lexical Access // Journal of Memory and Language 46, 245–266.

ОБ ОДНОМ СЛУЧАЕ НЕКАНОНЧЕСКОГО ИСПОЛЬЗОВАНИЯ МЕЖДОМЕТИЙ (КОРПУСНОЕ ИССЛЕДОВАНИЕ)[1]

Левонтина И. Б. (irina.levontina@mail.ru)
Институт русского языка им. В. В. Виноградова РАН

CORPUS-BASED STUDY OF NON-CANONICAL USE OF RUSSIAN INTERJECTIONS

Levontina I. B. (irina.levontina@mail.ru)
Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

The paper deals with the Russian interjections (*oj, oh, aj, ogo, uh*, etc.), namely their non-canonical use in collocations with K-words (Wh-words), mostly *kak* and *kakoj*. This type of use demonstrates a sort of syntactic recomposition — collocations *oj kak, oh kakoj*, etc. function as lexical units with the meaning of high degree, high quality or big quantity, although with very specific semantic shades. The paper makes use of the corpus data (the Russian National Corpus as well as the Internet data) to discover individual properties of interjections and their historical changes. Primary interjections are described against the background of interjections derived from the words of different part of speech. It turns out that in non-canonical use of primary interjections K-word can hardly be omitted, whereas derived interjections can also function the same way even without K-word. Non-canonical use of derived interjections is, with and without K-words, is very popular in contemporary Russian, especially in slang.

Key words: semantics, corpus-based study, interjections, non-canonical use, Russian language, syntactic recomposition, Wh-words

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

0. Введение

В настоящей работе рассматриваются неканонические употребления первичных междометий[2] в высказываниях типа:

(1) *После гибели родителей ему пришлось* ***ох*** *как тяжело.*

(2) *В разговоре с ребенком* ***ой*** *как непросто найти верный тон.*

(3) *Зарплата-то у нее* ***ого-го*** *какая!*

Подобные употребления резко отличаются от канонических контекстов с междометиями и семантически, и синтаксически; ср.:

(4) [Женя, Андрей Мягков, муж, 37, 1938] ***Ой*** */ какие мелкие кусочки!* ***Ой!*** [Эльдар Рязанов, Эмиль Брагинский. Ирония судьбы, или С легким паром, к/ф (1975)]

(5) ***Ого,*** *какая генеалогия интересная. Вот какие бабушки были.* [Женщина + мужчина: Брак (форум) (2004)]

В следующем примере присутствует и каноническое использование междометия *ах*, и неканоническое — междометия *ох*:

(6) ***Ах,*** *какая досада, надо было бы поступать на теплотехнический факультет, теперь бы знания* ***ой*** *как пригодились* [Анатолий Азольский. Лопушок // «Новый Мир», 1998]

Далее будет рассмотрена структура этого вида сочетаний и различия в поведении в них разных междометий; затем мы остановимся на их семантике, после чего поместим их в более широкий контекст и сравним поведение первичных и вторичных (производных) междометий. Исследование базируется на НКРЯ (ruscorpora.ru) с добавлением некоторого материала из личной переписки подростков вКонтакте.

1. Структура

При каноническом использовании междометия выступают в роли самостоятельных высказываний или в качестве вкраплений в высказывание, не имеющих синтаксических связей с ним [РГ-80 т. 1: 732-733]:

(7) *Ох!; Ох, нелегкая это работа (— из болота тащить бегемота).*

При описании синтаксиса междометий всегда отмечается, что некоторые из них обладают возможностью управлять падежными формами *(Увы мне, Ну тебя)*. Можно, однако, заметить, что нередко при междометиях

[2] Междометия — это класс неизменяемых слов, служащих для нерасчлененного выражения чувств, ощущений, душевных, состояний и других (часто непроизвольных) эмоциональных и эмоционально-волевых реакций на окружающую действительность [РГ 1980, т. 1: 732]. См. также [Шаронов 2008].

фигурируют относительные предложения, которые как бы заполняют валентность содержания:

(8) *Ох, какая нелегкая это работа!*

(9) *— Ого, какая погодка разыгралась, — заметил я, весело отряхиваясь от дождя в сторожке.* [А. К. Шеллер-Михайлов. Вешние грозы (1892)]

(10) *И деньга же перепадала — ух, какая деньга-то обильная!..* [В. В. Крестовский. Петербургские трущобы. (1864)]

(11) *Ай / какая красивая музыкальная шкатулка.* [Никита Михалков, Рустам Ибрагимбеков. Сибирский цирюльник, к/ф (1998)]

В таких случаях междометие факультативно, но без него местоимение выполняет уже не относительную функцию, а восклицательную:

(12) *Какая нелегкая это работа!*

Заметим, что таким образом могут использоваться совсем не все междометия; ср. неправильное

(13) **Ура, как удачно он выступил;*

(14) **Увы, как мало времени осталось.*

Разумеется, такое невозможно и для междометий с более специализированными значениями — например, для *эй*, основная функция которого состоит в привлечении внимания.

Далее фраза *Ох, нелегкая это работа* может подвергаться трансформации: происходит синтаксическое переразложение[3], в результате которого междометие «прирастает» к *к*-слову:

(15) *Работа это ох какая нелегкая!*

Интересно, что при этом междометие и *к*-слово почти утрачивают возможность использоваться друг без друга:

(16) *Работа это ?ох нелегкая!*

(17) *Работа это ??какая нелегкая!*

Оговоримся, что между каноническими и неканоническими контекстами существует большая область промежуточных употреблений междометий, которые можно трактовать по-разному:

---

[3] Понятие синтаксического переразложения не очень активно используется в лингвистике (в отличие от понятий переразложения и опрощения в морфемике), хотя явление синтаксического переразложения чрезвычайно важно, особенно в диахроническом аспекте. например, сейчас прямо на наших глазах набирает популярность оборот *Уже как (два дня)* — вместо *Уже два дня, как*. Тем не менее, это понятие иногда встречается в лингвистических работах. Так, в «Русской грамматике» (§ 1992), говорится об образовании сложных словосочетаний «на основе переразложения связей в предложении» [РГ-80 т. 2, 139].

(18) *И ох как они, страдальцы, обрадуются и воскреснут, откушав горячего.* [Виктор Астафьев. Затеси // «Новый Мир», 1999]

2. Избирательность

В рассматриваемой конструкции используется несколько первичных междометий, и ведут они себя в ней не одинаково.

Свободнее всего используются таким образом *ох* и в чуть меньшей степени *ой*[4].

2.1. Ох

(19) *Участие в таком деле, как моё, во всяком случае сделает его имя известным, а известность для адвоката,* ***ох****, как нужна!* [А. Ф. Кони. Из записок и воспоминаний судебного деятеля (1908)]

(20) *Мне его жизнь-то завидная, купеческая* ***ох*** *как горька была.* [А. Н. Арбузов. Таня (1938-1947)]

(21) *Когда человек бесталанный и при этом порядочный,* ***ох*** *как трудно ему, бедному, <…> ходить без дела.* [Константин Симонов. Так называемая личная жизнь/ Двадцать дней без войны (1973)]

Как видно из приводимых примеров, пунктуационное оформление подобных высказываний бывает двух типов В первом случае сочетание *ох как* никак не выделяется, как если бы это было обычное обстоятельство, во втором же *ох* выделяется запятыми, как если бы это было каноническое употребление междометия. Такая вариативность отражает промежуточный характер конструкции.

(22) *Почему уезжают умные, талантливые, серьезные люди, <…>, которые любят свою родину и* ***ох*** *как будут тосковать по ней?* [Виктор Некрасов. Кому это нужно? (1974)]

(23) *Сейчас я наткнулся— на предыдущей странице «Пакета» — ещё на один пример (а их в «Пакете»* ***ох*** *как много)* [Аркадий Мильчин. В лаборатории редактора Лидии Чуковской // «Октябрь», 2001]

(24) *А эта широкополая белая шелковая шляпа с бежевым кантом,* ***ох****, как ей дорога!* [Вацлав Михальский. Одинокому везде пустыня (2003)]

---

[4] Конечно, этот тезис уместно было бы подтвердить статистическими данными, однако сделать это затруднительно, во-первых, поскольку объем записей устной речи в НКРЯ пока не так велик, а рассматривая конструкция не очень частотна, а во-вторых, слишком большую долю материала составляют спорные или промежуточные случаи между каноническим и неканоническим типами употребления, что значительно снижает достоверность статистических данных.

В примерах (19-24) *ох как* можно заменить на *очень* — то есть, сочетание *ох как* реализует лексическую функцию Magn. В следующем примере ситуация чуть более сложная:

(25) *Уж я-то эту вашу публику* ***ох*** *как знаю.* [Филипп Янковский, Борис Акунин. Статский советник, к/ф (2005)]

Здесь перед нами не чистый Magn, а, возможно, отчасти и Bon.
Стоит остановиться на необычном и не вполне стандартном примере (26):

(26) *И, несмотря на то что актером, тем более популярным, становится,* ***ох****, как далеко не каждый, молодежь идет, валом валит «в кино», в артисты.* [Людмила Гурченко. Аплодисменты (1994-2003)]

Здесь *ох как* в значении высокой степени относится к *далеко не*, которое само по себе имеет значение высокой степени, поэтому пример нестандартный (ср., например, столь же сомнительное **очень весьма*).

Теперь перейдем к междометию *ой*.

2.2. Ой

(27) *…шел слух про этих бобылей, что из смирных становятся они —* ***ой*** *какие бойкие.* [А. Н. Толстой. Хождение по мукам/ Книга третья. Хмурое утро (1941)]

(28) *Работать с цветом в нашей кинематографии ещё* ***ой*** *как много надо!* [Григорий Горин. Чем открывается пиво? (1960-1985)]

(29) *Медведь хороший был, большой да жирный. Эти медведи* ***ой*** *какие лукавые!* [В. В. Верещагин. Из рассказов крестьянина-охотника (1895)]

(30) *это не так уж и плохо, лежать без ребенка — отдохнете, наберетесь сил, они вам дома* ***ой*** *как понадобятся, а наобщаться с ребенком вы успеете!* [Наши дети: Малыши до года (форум) (2004)]

(31) *…у меня там все мысли сводятся к суициду / поэтому я* ***ой*** *как не хочу влюбляться.* [Телефонный разговор московских студенток // Из коллекции НКРЯ, 2007]

Интересно, что в случае *ой*, как видно из приведенных примеров, предпочитается пунктуационное оформление по типу показателя степени, а не вводного слова.

Примеров на *ух* в корпусе меньше, но не так мало.

2.3. Ух

(32) *Тех же, кто мне не нравились, —* ***ух****, какими уродами я сделал!* [Г. Ф. Квитка-Основьяненко. Пан Халявский (1839)]

(33) *…сквозь любезность прокладывалась* ***ух*** *какая яркая прыть женского характера!* [Н. В. Гоголь. Мертвые души (1842)]

(34) *И хотя подчас в каждом приятном слове ее торчала* ***ух*** *какая булавка!* [Н. В. Гоголь. Мертвые души (1842)]

(35) *Мещанин он был, мой-то покойник, только,* ***ух****, какой ловкий, да умный, и начетчик был большой…* [А. А. Потехин. Виноватая (1868)]

(36) *— Двести тысяч — это* ***ух*** *какой капитал, а в хороших руках много с ним сделать дел можно.* [Н. Э. Гейнце. Выигрышный билет (1912)]

Следует отметить, что для ух такое употребление почти ушло, практически все примеры, которые удается найти в НКРЯ, старые.

2.4. Ах

(37) *Рассказывать неудобно, но вспомнить,* ***ах****, как приятно.* [Людмила Гурченко. Аплодисменты (1994-2003)]

(38) *Хотя, честно признаться, я не считала себя* ***ах*** *какой певицей — были студенты с голосами лучше моего.* [И. К. Архипова. Музыка жизни (1996)]

(39) *Не сказать, что дни эти были* ***ах*** *какими светлыми, но все-таки не такими, как день сегодняшний…* [Алексей Рыбин. Последняя игра (2000)]

Такие употребления ах нечасто, но встречаются, а вот, например, междометия *ай* и *эх* так не используются. Можно сказать *Ай, как неудобно получилось с Машей!* и *Эх, как досадно вышло!*, но не **С Машей получилось ай как неудобно* и не **Вышло эх как досадно*. Следует остановиться отдельно не междометии *ого*, однако это будет сделано в следующем разделе.

3. Разные стадии лексикализации

Естественно, что синтаксическое переразложение может сопровождаться лексикализацией. Рассмотрим некоторые связанные с этим явлением особенности отдельных сочетаний с междометиями. Так, особым образом распределены варианты *ого* и *огого*.

3.1. Ого(-го)

(40) *Одежда на нем почти сухая, а дождь шел* ***ого*** *какой!* [Н. Леонов, А. Макеев. Гроссмейстер сыска (2003)]

(41) *…болезни, особенно новые,* ***ого-го*** *каких бед могут натворить.* [Р. А. Сворень. Самое главное — понять самое главное // «Наука и жизнь», 2007]

(42) *О, дааа… тетеньки там* ***ого-го*** *какие))) как на подбор)*
[Переписка в icq между agd-ardin и Колючий друг (2008.01.16)]

(43) *Вавилов тоже* ***ого*** *какой умный был, я тебе точно говорю.*
[Александр Иличевский. Перс (2009)]

(44) *FireLove. Крошка ПУ никто сам по себе. А вот те, кто за ним стоят — это* ***ого*** *какие бабки…* [коллективный. Форум: Навальный (2012)]

(45) *Возможности-то у него* ***ого-го*** *какие!* [Александра Маринина. Последний рассвет (2013)]

Вариант *ого-го* тяготеет к лексикализации в функции интенсификатора. По-видимому, раньше гораздо активнее использовались и другие варианты междометий с редупликацией:

(46) *В Савинкове — да, есть что-то страшное. И* ***ой-ой****, какое трагичное.* [З. Н. Гиппиус. Дневники (1914-1928)]

Можно также обратить внимание на варианты *ух ты* и *ух*. Интересно, они распределены другим образом, чем *ого* и *ого-го*. Если у *ого* более длинный вариант *огого* предпочитается в немеждометном режиме, то у *ух* — более длинный вариант *ух ты* как раз в междометном (ср. неестественное **У него была ух ты какая зарплата*).

3.2. Ахти

Ахти в современном языке практически утратило способность к междометному употреблению и используется лишь в составе оборота *не ахти (как/какой)*:

(47) *Ну разумеется:* ***не ахти какая*** *радость об таком деле, да еще при людях толковать.* [А. В. Сухово-Кобылин. Дело (1861)]

(48) *— А сказать тебе по правде, — начала императрица ласковым тоном, — ведь и настоящий-то жених* ***не ахти какой****.* [Е. П. Карнович. Любовь и корона (1879)]

(49) *Признаться, и их воспитание было* ***не ахти каким*** *прочным: скорлупки внешнего, парадного благополучия очень легко лопались.* [К. С. Петров-Водкин. Моя повесть. (1932)]

(50) *Казалось бы, два года —* ***не ахти какой*** *долгий срок, а между тем много воды утекло за это время.* [Е. А. Аверьянова. Иринкино счастье (1910)]

(51) *Конечно,* ***не ахти*** *какой вкус был у Кастанье.* [Ю. О. Домбровский. Хранитель древностей, часть 1 (1964)]

(52) *И вообще пора закругляться с этим. Деньги* ***не ахти какие****, а репутация… Елтышев вышел.* [Роман Сенчин. Елтышевы (2008) // «Дружба Народов», 2009]

(53) *Ноябрь перевалил на вторую половину, погода* ***не ахти****, насладиться природой и свежим воздухом все равно не получится.* [Александра Маринина. Последний рассвет (2013)]

В последнем примере *не ахти* используется в значении невысокой степени самостоятельно, без управляющего слова (то есть, без *к*-слова). Такое употребление присуще и другим степенным словам; ср. *Погода не очень, Фильм не особенно.*

Раньше, однако, *ахти* могло употребляться как междометие:

(54) *Но **ахти**! продолжал я: где-то он ныне?* [А. Т. Болотов. Жизнь и приключения Андрея Болотова, описанные самим им для своих потомков (1800)]

(55) *— **Ахти**, матушка! — вскричала старостиха, всплеснув руками.* [Д. В. Григорович. Бобыль (1847)]

(56) *— Да о вашем выступлении третьего дня. Забыли? — **Ахти**! Вам уже донесли? — А как же.* [И. Грекова. На испытаниях (1967)]

Сейчас эти примеры выглядят совершенно устаревшими, а последний, по-видимому, стилизованный.[5]

3.3. Повторы

Наконец, стоит отметить, что для с неканонического использования междометий чрезвычайно характерна конструкция с повтором вида *X, ох/ой как/какой X*:

(57) *…прошли праздничные, медовые дни прекраснодушия и наступили суровые (**ой**, какие суровые!)* [З. Н. Гиппиус. Дневники (1914-1928)]

(58) *Этот господин, который похож на вас… он очень умный, **ух**, какой умный…* [В. Г. Короленко. Братья Мендель (1915)]

(59) *Трудным, **ой** каким трудным стал для нее XX век.* [«Жизнь национальностей», 2000.06.23]

(60) *Обидно, **ах**, как обидно. И та же самая обида гнетет всех нас — всех женщин России.* [Е. Пищикова. Город после мифа // «Русская жизнь», 2012]

4. Семантика

Рассмотренные сочетания имеют значение высокой степени, положительной или отрицательной оценки или большого размера/количества (последнее особенно для *ого-го)*. Интересно при этом, что благодаря междометию возникает эффект цитирования — как бы контрабанда эгоцентрического в нарративе:

(61) *Рассказывая <…> бабушка очень оживлялась и в паузах многозначительно кивала сама себе головой, как бы давая понять, что ещё **ой** как много она бы понарассказала, но не будет.* [Марина Палей. Поминовение (1987)]

---

[5] Междометие *охти* также существовало, но практически не сохранилось.

(62) *Как-то одна из дам, гостивших у Ляли, воскликнула: «**Ах**, какая у вас кофточка!..» И ушла домой в этой «**ах**, какой кофточке».* [Евгений Весник. Дарю, что помню (1997)]

Последний пример особенно показателен — в нем представлено прямое цитирование восклицания и демонстрируется механизм перехода от канонического использования междометия к неканоническому.

Естественно, что при неканоническом употреблении происходит десемантизация междометий, и смысловые различия между ними частично нейтрализуются (как, например, различия *ох* и *ой*). Тем не менее, полностью разные междометия не синонимизируются. Ср. следующий пример, в котором едва ли можно поменять *ах* и *ух* местами:

(63) *Добавьте сюда также мнения жен, любовниц и случайных знакомых, которые тоже **ах как** любят киношную романтику и крутость, и которых **ух как** раздражает вечно топчущийся за дверью «посторонний».* [Сергей Козлов. Волшебники-недоучки (2004) // «Боевое искусство планеты», 2004.03.11]

5. Более широкий контекст

Рассмотренные образования с междометиями вписываются в более широкий круг явлений.

5.1. Лексикализация сочетаний с *к*-словами

С одной стороны, в русском языке есть большая группа единиц разной степени лексикализованности, образованных на основе сочетаний *к*-слов с компонентами самой разной природы. Это, в частности, неопределенные местоимения на *-то, -либо, -нибудь, -кое*, а также близкие к ним сочетания на *угодно, попало, попадя, придется, абы, придется, хочешь, Бог на душу положит, черт знает, не пойми-разбери* и т. д. См. о некоторых из этих сочетаний [Левонтина, Шмелев 2005].

Можно также отметить лексикализованное сочетание указательной частицы *во* с *к*-словами (обычно с *как*), указывающее на столь высокую степень какой-то характеристики или потребности, что ее трудно или невозможно выносить:

(64) *У меня мама болеет, братья еще не приехали. А мне в Иваново **во как** надо! — И я ребром ладони провел по горлу.* [Вальтер Запашный. Риск. Борьба. Любовь (1998-2004)]

В речи выражение *во как* в подобных контекстах часто сопровождается характерным жестом — говорящий подносит ладонь горизонтально к горлу (часто как бы обхватывая горло большим и указательным пальцем) или делает ладонью движение, как бы перерезая себе горло.

(65) — *А я, понимаешь, пять лет по Соссюру лингвистику читал, Леви-Стросса, Якобсона, Бахтина, как братишек, люблю и —* ***во как*** *уважаю!* [Александр Иличевский. Бутылка (2005) // «Зарубежные записки», 2008]

(66) *…диваны заказчику* ***во как*** *нужны! Но его кошелек уже пуст.* [«Мебельный бизнес», 2003.05.15]

(67) *Мне наши музейные дела* ***во как*** *горло переели, ну вот и сговорился я с молоденькой сотрудницей и поехал с ней в выходной.* [Ю. О. Домбровский. Факультет ненужных вещей, часть 5 (1978)]

С другой стороны, имеет смысл обратиться и к вторичным междометиям (*ужас, смерть* и т. п.), которые во многом ведут себя подобно первичным.

5.2. Смерть, ужас

Приведем пример канонического использования вторичного междометия *смерть*:

(68) *И шлейфом шелесть / шелесть / шелесть / шелесть / шелесть!* ***Смерть*** *как я влюблена!* [Виктор Иванов, Михаил Старицкий. За двумя зайцами, к/ф (1961)] [аналогично — *Ах, как я влюблена*].

Однако может происходить и переразложение, при котором *смерть* прирастает к *к*-слову:

(69) *А что говорят в народе / будто мы чокнутые / так это поклёп. Нам это просто* ***смерть*** *как интересно.* [Алла Сурикова и др. Чокнутые, к/ф (1991)]

Неканонические употребления особенно характерны для вторичного междометия *ужас*:

(70) *Это здорово! Вы* ***ужас*** *какие молодцы! Мы все думаем о вас и гордимся нашим Ленинградом.* [Дмитрий Каралис. Мы строим дом (1987-2001)]

(71) *Мы схватили эту трубу и понесли. Но она оказалась* ***ужас*** *какая тяжёлая!* [Ирина Пивоварова. Верная собака Уран (2001)]

(72) *Ниловна идеи сына приняла, а сынок ведь,* ***ужас*** *каким асоциальным делом занимался -).* [Наши дети: Подростки (2004)]

(73) *Почему-то когда речь заходит о его Кире, Паша становится* ***ужас*** *каким грубым.* [Дина Сабитова. Где нет зимы (2011)]

Аналогично ведут себя междометия страх и страсть (заметим, что междометие страсть восходит к старому значению существительного страсть — что-то страшное):

5.3. Страх/страсть

(74) *Прохарчин <…> **страх** как полюбил отыскивать истину.* [Ф. М. Достоевский. Господин Прохарчин (1846)]

(75) ***Страсть** как она, стерва, сладенькое любит…* [Ю. О. Домбровский. Факультет ненужных вещей (1978)]

(76) *Рядом со мной стоял мальчик, мне **страх** как хотелось с ним поговорить.* [И. Грекова. В вагоне (1983)]

(77) *…все девушки романы потихоньку читали. Один раз мне попался **страх** какой интересный.* [З. Н. Гиппиус. Простая жизнь (1896)]

(78) *Сам Великий артист давно уже не принимал спиртного, угощать же как истинный хлебосольный сибиряк **страсть** как любит.* [Виктор Астафьев. Затеси (1999) // «Новый Мир», 2000]

(79) *Кровища хлещет, а нам-то **страх** как весело.* [Евгения Пищикова. Пятиэтажная Россия (2007) // «Русская Жизнь», 2008]

(80) *— Ивану **страсть** как неловко было, что деньги— где-то у жены «в чулочке»* [Василий Шукшин. Печки-лавочки (1970-1972)]

С точки зрения возможности неканонического употребления вторичные междометия также очень избирательны. Например, *жуть* тоже так используется, а *кошмар* — едва ли.

(81) *Конечно, мне **жуть** как интересно поглядеть на того, кто так рьяно добивается моей руки.* [Андрей Белянин. Свирепый ландграф (1999)]

(82) *:) …будут с пеной у рта доказывать ей, что она ну **жуть** как несчастна.* [Женщина + мужчина: Брак (форум) (2004)]

Ср. сомнительное ??*Мне кошмар как не хочется туда ехать.*

5.4. Неканоническое использование междометий в современном сленге

В современном разговорном языке, особенно в сленге, такая модель очень продуктивна:

(83) *а главное, это было **трендец** как своевременно*

(84) *ей **охренеть** как нужны эти отношения*

(85) *Просто **капец** какая злая*

(86) *А я еще и **кАпец** как быстро бегаю*

(87) *А я залипаю **п*пец** как на этой песне*

(88) *Ну ей страшно как **песец***

5.5. Различие между конструкциями с первичными и вторичными междометиями

И вот здесь обнаруживается существенное различие между конструкциями с короткими междометиями и другими частями речи в аналогичной функции. Первые практически не могут использоваться без *к*-слов. Последние легко начинают использоваться в такой роли самостоятельно:

(89) ***Страх*** *люблю!*

(90) ***Смерть*** *люблю!*

(91) *Это было* ***охренеть*** *вовремя!*

(92) *Ты* ***п*пец*** *права*

(93) *Просто* ***я п*пец*** *не хочу чтобы знала вся Россия*

(94) *П*пец, его мнение* ***п*пец*** *много значит*

(95) *Еще ты* ***п*пец*** *красотка*

(96) *Ну мне реал* ***п*пец*** *хреново*

(97) *У меня прост* ***п*пец*** *плохое настроение*

(98) ***Капец*** *она вам задаёт …*

(99) ***Капец*** *я отстойная, да?* 😊

(100) *Это вообще* ***капец*** *обидно*

References

1. *Levontina I. B., Shmelev A. D.* (2005), Poorly explored items with the meaning of ‘not specified choice criteria’ in Russian [O nekototyh maloizuchennyh jedinicah so znachenijem nezadannosti kriterijev vybora v russkom jazyke], Logical analysis of language: quantitative aspect of language, Indrik, Moscow, p. 638–651.
2. *Russian Grammar* (1980), V. 2, Moscow.
3. *Sharonov I. A.* (2008), Interjections in Speech, Language and Dictionary [Mezhdometija v rechi, tekste i slovare], Moscow.

АБЫ: КОРПУСНОЕ ИССЛЕДОВАНИЕ В АСПЕКТЕ СИНХРОНИИ И ДИАХРОНИИ[1]

Левонтина И. Б. (irina.levontina@mail.ru)
Институт русского языка им. В. В. Виноградова РАН

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

THE RUSSIAN *ABY*: CORPUS-DRIVEN RESEARCH (SYNCHRONY AND DIACHRONY)

Levontina I. B. (irina.levontina@mail.ru)
Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

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

The paper deals with the Russian *aby* as a marker of "free choice" (or, rather, not specified choice criteria) within indefinite pronouns against the background of other markers of "free choice" such as *ugodno*, *popalo*, *pridetsia*. It pays attention not only to the synchronic semantics of *aby*, but also to its history and claims that the modern meaning of *aby* is related to its usage as a conjunction. The paper makes use of the corpus data (the Russian National Corpus as well as the Internet data) to follow the changes in the use of the particle in question over the last two hundred years. It investigates into the range of *K*-words that can collocate with *aby*: the most typical are collocations with *kto*, *chto*, *kak* and *kakoi*; however, collocations with other *K*-words are also present in the corpora. In addition, it discusses the question of negative polarity of *aby* and the increasing degree of its polarization.

Key words: semantics, pronouns, "free choice", linguistic change, diachronic corpus research

[1] Работа выполнена в рамках научно-исследовательского проекта РФФИ № 16-06-00339 «Контрастивное корпусное исследование дискурсивных слов русского языка».

1. Незаданность критериев выбора и местоимения с *абы*

В самом начале 2000х внимание сразу нескольких исследователей было привлечено к русским местоименным единицам со значением незаданности критериев выбора референта — таким, как *что угодно*, *что попало*, *что придется* (ср., в частности, статью [Левонтина, Шмелев 2005], в основу которой положен доклад авторов на конференции «Логический анализ языка» в июне 2004, и доклад [Тестелец, Былинина 2005], прочитанный в ИППИ в феврале 2005). Был описан целый ряд нетривиальных языковых свойств рассматриваемых единиц, некоторые из которых связаны с тем, что элементы, входящие в их состав, оторвались от полнозначных единиц, к которым они восходят, но в то же время сохраняют память о своем происхождении. Многие элементы, входящие в состав серийных неопределенных местоимений, включают в свой состав элементы, восходящие к полнозначным словам и словосочетаниям (*либо*, *нибудь*); однако всякая связь с исходными единицами (*любо*, *ни будь*) в современном языке полностью утрачена. Напротив того, для единиц типа *угодно*, *попало*, *придется* эта связь нередко осознается говорящими и в каких-то случаях может оказаться актуальной или же актуализоваться[2]. Если в сочетании *где угодно* (и в большинстве подобных сочетаний) от желательности практически ничего не остается, то в выражении *сколько угодно* в каком-то виде связь с желанием просвечивает; ср.:

(1) *Спустись в подвал — там сколько угодно вина.*

(2) **Не спускайся в подвал — там сколько угодно крыс.*

Правильность первой фразы и неправильность второй объясняется тем, что в случае с вином речь идет о чем-то желательном, а в случае с крысами — о чем-то неприятном или даже опасном [Левонтина, Шмелев 2005: 647].

При этом, хотя вследствие идиоматизации значение рассматриваемых единиц в значительной степени выхолащивается, между ними сохраняются различия. Ср. следующий пример, в котором нельзя было бы поменять местами обороты *кто угодно* и *черт знает кто*:

(3) *… «Спартак» — это та команда, которая способна как обыграть кого угодно, так и продуть черт знает кому…* [Ю. Гладильщиков]

Пути идиоматизации таких местоименных единиц удобно проследить на основе данных лингвистических корпусов, охватывающих более или менее продолжительный период времени (достаточный для того, чтобы языковые изменения были наблюдаемы). В настоящей статье речь пойдет о частице *абы* в составе местоименных выражений, которой до сих пор почти не уделялось внимания со стороны исследователей.

---

[2] В зависимости от того, насколько указанная связь остается актуальной, такие выражения, как *кто угодно*, *где придется*, *как попало* и т. п., могут описываться как свободные сочетания, как идиомы или как полноценные местоимения.

Привычным является такое использование корпусных данных, при котором выдвигаются те или иные гипотезы, которые в дальнейшем проверяются на основе статистического анализа. Однако не менее плодотворным может оказаться подход, при котором сами гипотезы формулируются на основе изучения большого массива корпусных данных. Разумеется, эти два подхода не противоречат друг другу.

Заметим, что в «Малом академическом словаре» [Евгеньева 1981] для слова *абы* (включенного в словарь с пометой *Обл.*) указано лишь союзное употребление (с толкованием «Лишь бы, только бы» и иллюстрацией из «Тихого Дона») и лишь за знаком ромба в качестве фраземы приводится выражение *абы как* с пометой *прост.*, толкованием «как-нибудь» и иллюстрацией из «Поднятой целины». Эта словарная статья практически без изменений (единственное изменение состоит в том, что в новом издании допущено двоякое ударение) перенесена и в новое издание «Малого академического словаря» [Крысин 2016].

На фоне других единиц со значением незаданности критериев выбора *абы* обладает ярким семантическим обликом. Если фраза *Ест что попало* указывает на неразборчивость субъекта, *Ест что придется* — на его неприхотливость (причем в обоих случаях субъект не обязательно стремится поесть во что бы то ни стало), а фраза *Ест что угодно* подразумевает, что субъекту можно предложить даже самые странные кушанья[3] (подробнее об особенностях семантики оборотов с *попало*, *придется* и *угодно* см. [Левонтина, Шмелев 2005]), то *Ест абы что* предполагает, что субъект непременно что-нибудь ест, но ему все равно что.

Ср. характерный пример:

(4) *Муза ко мне так и не собралась. Нужно идти сдаваться. Приходить пустым рискованно, я накидал* ***абы каких*** *глупостей. Кетчуп с овощами можно назвать «Дача». Будто бы из помидоров, выращенных на частных грядках.* [Слава Сэ. Ева (2010)]

Здесь прямо сказано, что человек пытается принести все равно что, но принести хоть что-нибудь, потому что приходить с пустыми руками неудобно.

Сходным образом устроены следующие три примера:

(5) *Не верь ей. Просто с возрастом приперло. Решила родить* ***абы от кого*** *— ну, а тут как раз дурак попался. Никого она, кроме себя, не любит, и никогда не любила…* [Вячеслав Рыбаков. Гравилет «Цесаревич» (1993)]

(6) *Гаранин и Настя вошли в квартиру, тесную, двухкомнатную, с дурной, непродуманной планировкой, словно строили ее наспех,* ***абы как****, лишь бы отвязаться.* [Влада Валеева. Скорая помощь (2002)]

(7) *Главная беда заключалась в том, что многие мастерские шили по старинке:* ***абы как****, все сойдет, все возьмут.* [Виктор Жизнев. Одежда, в которой приятно работать (2003) // «Встреча» (Дубна), 2003.03.26]

[3] Заметим, что этот пример иллюстрирует высказанное выше положение, согласно которому в большинстве случаев сочетания с *угодно* утрачивают компонент желательности.

2. Значение и употребление *абы* в современном русском языке по корпусным данным

2.1. Негативная полярность

Яркая черта единицы *абы* — тяготение к негативной полярности[4]. Весьма часто *абы* используется в противительной конструкции вида «*не абы кслово, а…*» (из 196 вхождений оборотов вида «*абы кслово*» в основном корпусе НКРЯ по состоянию на 20 февраля 2018 целых 41 — это именно конструкции вида «*не абы кслово, а…*»):

(8) *Он любил читать, да* ***не абы что, а*** *стихи и философскую литературу.* [Преферанс его жизни (2003) // «Криминальная хроника», 2003.06.10]

(9) *В ней на голубом глазу излагается кодекс поведения девушки, которая хочет выйти замуж, но* ***не*** *просто* ***за абы кого, а*** *за того, кто её действительно любит и, главное, будет любить всю оставшуюся жизнь.* [Ксения Махненко. Обращение (2002) // «Домовой», 2002.03.04]

(10) *А так как дядя Витя начал свой путь с ПТУ, то решено было не искушать судьбу, а идти в это учебное заведение, и* ***не абы в какое, а*** *именно в то же самое — ПТУ номер один при заводе ЗИЛ.* [Алексей Моторов. Преступление доктора Паровозова (2013)]

(11) *Это при том, что у нас были хорошие учителя, которые стремились что-то до нас донести, и* ***не абы как, а*** *заинтересовать нас.* [коллективный. Форум: Школьные рамки (2013)]

Впрочем, наличие эксплицитного противопоставления не обязательно (в основном подкорпусе НКРЯ по состоянию на 20 февраля 2018 находим еще 39 примеров с отрицанием *не* при отсутствии эксплицитного противопоставления):

(12) *Ее снимала Энни Лейбовиц, а Энни Лейбовиц* ***абы кого не снимает****.* [Игорь Найденов, Наталья Водянова. Личное тело Натальи Водяновой // «Русский репортер», 2012]

(13) *Если ресторан попроще, то и женщина будет попроще, — безжалостно оскалился Воск. — Марину* ***абы куда не поведешь****…* [Даниил Корецкий. Менты не ангелы, но… (2011)]

(14) *Ты что, монашка? Мужа ведь не нашла, это я и без расспросов поняла, по тебе сразу видно. — Не нашла. Но* ***абы кого*** *мне тоже* ***не нужно****.* [Вацлав Михальский. Прощеное воскресенье // Октябрь, 2009]

Нередко *абы* фигурирует в негативно поляризованных контекстах, организованных более сложным образом:

[4] Это свойство отмечалось в работах [Левонтина, Шмелев 2005; Тестелец, Былинина 2005]

(15) *С другой стороны,* ***не хотелось*** *и отдавать квартиреху* ***абы как****, за бесценок.* [Андрей Волос. Недвижимость (2000) // «Новый Мир», 2001]

(16) *Учиться* ***абы как*** *она* ***не умела*** *и мерно тянула воз.* [Сергей Каледин. Аллея Руж // «Огонек», 2013]

(17) *я думаю,* ***абы кому*** *проводить Олимпиаду и ЧМ по футболу* ***не доверят****…* [коллективный. Форум: Жители Пугачева из-за убийства десантника перекрывали трассу (2013)]

(18) *Он поступал в аспирантуру, его профессор лично звал —* ***не думаю****, что звали бы* ***абы кого****!* [коллективный. Форум: Большая перемена (2001–2011)]

Иногда отрицание скрыто в лексическом значении одного из слов, в контексте которого появляется *абы*:

(19) *Все звали в рестораны, норовили оказать какие-то услуги… Но Петрицкий* ***запретил*** *дружить* ***абы с кем****.* [Даниил Корецкий. Менты не ангелы, но… (2011)]

В вышеприведенном примере (18) *не думаю* можно было бы заменить на *сомневаюсь*, и употребление *абы* было бы столь же естественным — ср. (20):

(20) *Он поступал в аспирантуру, его профессор лично звал —* ***сомневаюсь****, что звали бы* ***абы кого****!*

Однако, как показывает корпусный материал, примеры, в которых отсутствует отрицательная поляризация встречаются не столь уж редко.

Попадаются высказывания, в которых подчеркнута идея случайности выбора, и это оказывается достаточным основанием для употребления *абы*:

(21) *…я определенно за форму. Ибо это придает, ну, что-то свое каждой школе. Когда я училась в начальной, у нас была школьная форма, конечно не фонтан, но была. Было оригинально, ибо все остальные школы ходили* ***абы как****, а мы определенно. Все знали из какой школы ребенок, и это было интересно.* [коллективный. Форум: Школьная форма. За и против (2007–2010)]

Для автора записи даже плохая форма (*не фонтан*) лучше, чем произвольно выбранная одежда.

Но чаще встречаются примеры, в которых идея рандомности сопровождается идеей низкого качества или небрежности:

(22) *Пригласили крышу перекрывать неверующих, они запили, неделю делали и сделали* ***абы как****, а тут — один день и высочайшее качество.* [В. Н. Павленко, К. Ваннер. Особенности психологии евангельских христиан-баптистов (2004) // «Вопросы психологии», 2004.10.12]

(23) *Спросить не с кого, зато и свое дело можно сделать* ***абы как****,* ***тяп-ляп****.* [Время события люди (2003) // «Встреча» (Дубна), 2003.02.12]

(24) *Да и соблюдался-то он в те недолгие годы* ***абы как, спустя рукава****.* [Алла Лерчик. Жемчуг слёз и розы смеха // «Зеркало мира», 2012]

(25) *— Олег Николаевич открыл другую бутылку коньяка. — Учится —* ***абы как! Работать не желает!*** [Карен Шахназаров. Курьер (1986)]

Даже в тех случаях, когда идея низкого качества не профилирована, читатель часто бывает склонен вычитывать эту идею из текста:

(26) *От главной улицы вверх, в сторону бора, — частые переулки, там уже и ворот нет, и ограды далеко не везде, городьба поставлена* ***абы как*** *— и плетень, и жердянник, и просто подсолнухи посажены полосой погуще, вот тебе и грань между дворами.* [Сергей Залыгин. Соленая Падь (1967)]

Строго говоря, речь идет только о случайном выборе средств разметки, однако возникает ощущение неопрятности и запустения.

То, что идея рандомности легко провоцирует отрицательную оценку, присутствует и в сочетаниях с *попало*. Не случайно *абы* иногда встречается в одной фразе с *попало*:

(27) *Как будто вдохновение делало его чужим в своей квартире, случайным гостем, стесняющимся занимать место хозяина и пристраивающимся* ***абы как где попало****.* [Евгений Чижов. Перевод с подстрочника (2012)]

Между прочим, между *абы* и *попало* обнаруживается интересное различие в принадлежности отрицательной оценки. Если *попало* допускает оценку со стороны, так что фраза *Вечно ты дружишь с кем попало* имеет два понимания: либо говорящий не одобряет друзей, либо подразумевает, что адресат не задумывается, с кем дружить, — то фраза *Вечно ты дружишь абы с кем* имеет только одно понимание: адресату непременно нужны друзья, все равно какие.

2.2. Избирательность в отношении *к-слов*

В подавляющем большинстве примеров *абы* сочетается со словами *кто*, *что*, *как* и *какой*:

(28) *Но если Кира была готова бежать под венец* ***абы с кем****, то Катя хотела большой любви.* [Дарья Донцова. Доллары царя Гороха (2004)]

(29) *Ну еще в студию приходят гости, и не* ***абы кто*** *с улицы, а именитые актеры, политики, певцы и другие деятели шоу-бизнеса.* [коллективный. Форум: Комментарии к передаче «ПрожекторПерисХилтон» (2009–2011)]

(30) *И у нас не принято ходить на работу* ***абы в чем****.* [А. В. Жвалевский, Е. Пастернак. Время всегда хорошее (2009)]

(31) *Конечно, рисовать надо не* ***абы что****, а нечто приемлемое для данной команды, и не* ***абы как****, а в том стиле, в каком делаются и другие надписи*

этой группы. [Илья Абель. Новый палимпсест или краткий словарь иностранных слов // «Знание — сила», 2005]

(32) *Всех наша мамочка учила относиться к делу не* ***абы как****, а с полной ответственностью.* [коллективный. Сердце матери (2013.03.15) // «Новгородские ведомости», 2013]

(33) *Да ему нужна не* ***абы какая****, а чтобы и богатая, и с квартирой, и чтобы ему трусы покупала.* [Маша Трауб. Ласточ…ка (2012)]

Другие сочетания в НКРЯ представлены плохо или совсем не представлены, однако в эл. интернете присутствуют. Нередко встречаются примеры на *абы когда* и *абы где*:

(34) *Ну, комиссию* ***абы когда*** *не соберёшь, нужен приказ по вузу и нужно, чтобы члены комиссии могли присутствовать…*

(35) *В центре деревни стоит деревянная протестанская церковь. Вокруг нее кладбище. Построена не* ***абы когда****, а в 1667 году!*

(36) *Мальчик, названный Андрюшей в честь героического папы, появился на свет не* ***абы когда****, а в самый день победы — 9 мая 1945 года.*

(37) ***Абы где*** *компьютер лучше не покупать. Шоппинг в интернете становится все популярнее. Покупки в интернете могут быть еще более выгодными.*

(38) *При этом квартиру не* ***абы где****, а в Москве, желательно в хорошем районе.*

Более того, попадаются примеры на *абы куда* и *абы откуда*:

(39) *Так все сразу бесятся-как так?? иди* ***абы куда*** *работать? я ищу исключительно вакансию по редкой специальности.*

(40) *Но если ее нет, то надо ее сначала подготовить и создавать, чтобы музеи не переселялись* ***абы куда****, в какие-то мало приспособленные для этого здания.*

(41) *Мурамаса в филлерах — он же не* ***абы откуда*** *взялся! Момото Мурамаса — это древний изготовитель мечей в Японии…*

(42) *Вот прям никогда не логиньтесь* ***абы откуда****.*

Как ни странно, есть даже сочетания *абы зачем* и *абы почему*:

(43) *Дураку понятно, что столь высокие фигуры* ***абы зачем*** *и абы к кому не ездят…*

(44) *…Царевна-лягушка является в русских сказках не* ***абы почему****, а потому…*

Отдельно стоит отметить сочетание *абы сколько*:

(45) *…в наше время деньги нужны всем и все хотят их заработать и не* ***абы сколько****, а чтобы хватало.*

Впрочем, по какой-то причине именно на сочетание *абы сколько* в интернете встречается особенно много нестандартных примеров, в которых более обычным было бы использование, например, сочетания *черт-те сколько*:

(46) *Сегодня простояла в очереди на почте* ***абы сколько*** *времени всего за тремя конвертиками с марками.*

(47) *Он, мой супруг, некогда выбранный мною в качестве спутника жизни, перестал сворачиваться в кокон на диване после работы, и дрыхнуть по вечерам* ***абы сколько****.*

(48) *Мазь, пылившаяся на полке* ***абы сколько*** *лет, имела свой срок годности, но родители, видимо, считали, что пока она не растворила тюбик…*

В примерах такого рода у сочетания *абы сколько* появляется импликатура ‘много’. Это сходно с появлением точно такой же импликатуры у сочетания *сколько угодно* (см. подробнее [Левонтина, Шмелев 2005: 645–646]).

Таким образом, хотя *абы* в своей сочетаемости имеет определенные предпочтения, в целом оно способно сочетаться практически с любыми ксловами.

2.3. Союзные употребления *абы*

Такие употребления составляют незначительную часть современного корпусного материала, а кроме того, они в основном устарели: из 127 вхождений *абы* в созданные в 2000 или позже тексты основного подкорпуса НКРЯ (по состоянию на 20 февраля 2018) к союзным употреблениям относятся лишь шесть вхождений, одно из которых — из исторического романа Александра Архангельского «Александр I» — представляет собою явную стилизацию. В большинстве случаев союзное *абы* может быть заменено на нейтральное *лишь бы*. Ср. примеры из «Национального корпуса русского языка» (НКРЯ):

(49) *Ну, а там, кто знает: может «в вышнем суждено совете», что я должен овладеть начатками всех специальностей, какие есть в лагере,* ***абы*** *потруднее…* [Юлий Даниэль. Письма из заключения (1966–1970)]

(50) *Когда пары сводят случайно, наугад,* ***абы*** *взлетели,* ***абы*** *поднялись и ушли, летчики, не зная друг друга, ярятся и психуют…* [Артем Анфиногенов. А внизу была земля (1982)]

(51) *На то, что вас будут встречать с хоругвями, и не надейтесь, милая.* ***Абы*** *живой остаться.* [В. Лихоносов. Ненаписанные воспоминания. Наш маленький Париж. Ч. 3–4 (1983)]

(52) *Но, думаешь, хотя бы половина клиентов понимает, что это такое? — А зачем им понимать? — пожала плечами Линка. —* ***Абы*** *деньги платили.* [Анна Берсенева. Полет над разлукой (2003–2005)]

Эти примеры показывают, что тот смысловой компонент, который отличает *абы* от *угодно*, *попало* и *придется* и о котором шла речь выше, восходит

именно к союзному употреблению *абы*. В этом отношении очень показателен следующий пример:

(53) *Тиша решил отойти от общественной жизни, подался на лесосплав в Саяны,* ***абы куда подальше****, старался забыть про все, что случилось в Тетеревке, но выковырять чувство вины не мог.*
[Евгений Евтушенко. Ягодные места (1982)]

Здесь в выделенном обороте *абы* употреблено в союзном значении, однако ясно виден механизм перехода от союзного употребления к «кванторному».

Следует отметить, что в НКРЯ также представлено некоторое количество примеров, в которых союз *абы* используется в значениях, отсутствующих в современном литературном языке:

(54) *Их дело. Молодые.* ***Абы*** *для здоровья полезно.*
[Михаил Анчаров. Как Птица Гаруда (1989)]

(55) *«Кисленького бы чего сейчас* ***абы*** *солененького, — непроизвольно вырвалось, дурнота вдруг подкатила к горлу и тут же пропала.*
[Владимир Личутин. Любостай (1987)]

3. Изменение употребления

Материал НКРЯ показывает, что употребление *абы* сильно изменилось за последние 200 лет. До начала XX в. оно используется почти исключительно как союз, употребления с кусловами активизируются лишь с середины XX в., а тенденция к преобладанию негативной полярности совсем поздняя. Во всех без исключения примерах XIX в. из основного подкорпуса НКРЯ (всего 120 вхождений) представлено союзное употребление *абы*:

(56) *— Государь Ива Олелькович, невеста твоя едет, иди навстречу,* ***абы*** *хищник Кощей не исхитил ее!* [А. Ф. Вельтман. Кощей бессмертный. Былина старого времени (1833)]

(57) *Находили, однакоже, все нужное по дороге, и Кузьма всегда приговаривал: «****Абы*** *гроши — все будет».* [Г. Ф. Квитка-Основьяненко. Пан Халявский (1839)]

(58) *Сестра его страсть как боялась, а мать хоть и не боялась, но часто по его делала, «****абы*** *лихо спало тихо».* [Н. С. Лесков. Житие одной бабы (1863)]

(59) *Нам все равно, что фригийский колпак, что Мономахова шапка, —* ***абы*** *мы были целы.* [Н. С. Лесков. Божедомы (1868)]

(60) *Лучше же умыть руки, сделать достодолжное и — дальнейшее предоставить администрации: как там себе хотят, так пусть и делают,* ***абы*** *мы в стороне были!* [В. В. Крестовский. Панургово стадо (Ч. 1–2) (1869)]

(61) *Пустое это и не господское дело лошадьми торговать, но, думаю, чем бы дитя ни тешилось,* ***абы*** *не плакало, и говорю: «Извольте».*
[Н. С. Лесков. Очарованный странник (1873)]

(62) *Ноне не токма что застоять, а потопить норовит всякий… **абы** самому сухому из воды выбраться…* [А. И. Эртель. Записки Степняка (1883)]

Этот тип употребления хорошо сохраняется и в первой половине XX в.:

(63) *И в том тебе каюсь и душевно пишу, **абы** на меня по напраслине ты не гневался.* [Ал. П. Чехов. Письма Антону Павловичу Чехову (1902)]

(64) *Подумаешь… **Абы** жизнь была — богов выдумают… —* [Вс. В. Иванов. Бронепоезд № 14.69 (1922)]

(65) *Это от бедности своей она совершенно инертна в общественных начинаниях, потому что бедный человек думает только «**абы** просуществовать» и начинание считает роскошью.* [М. М. Пришвин. Дневники (1924)]

(66) *За чужим не гонюся, **абы** свое было цело!* [Л. М. Леонов. Вор. Часть 3 (1927)]

(67) *— Я разобьюсь на части, а все сделаю, **абы** она со мной поехала…* [А. С. Макаренко. Педагогическая поэма. Часть 2 (1934)]

Но уже с начала XX в. появляется и постепенно нарастает приместоименное употребление *абы*:

(68) *И все это он знал не **абы как**, а вполне основательно и надежно.* [П. И. Ковалевский. Петр Великий и его гений (1900–1910)]

(69) *И будто умные люди, из панов, а так **абы что** говорят.* [А. С. Серафимович. Бомбы (1906)]

(70) *Вижу я, да и другие тоже не слепые, что Феня эта моя такая, выходит, птица, которая по части гнезда вполне вить умеющая, и даже с большою она жадностью к этому делу, а не то как другие бывают — **абы как**!* [С. Н. Сергеев-Ценский. Кость в голове (1932)]

(71) *В лавках нонешний год их не было… **Абы в чем** лето проходил… В церкву ажник страмно идтить в старой.* [М. А. Шолохов. Тихий Дон. Книга вторая (1928–1940)]

(72) *Делай **абы как**. Чего ты стараешься?* [М. А. Шолохов. Тихий Дон. Книга четвёртая (1928–1940)]

(73) ***Абы кого** с такими предосторожностями этапировать не станут!* [Г. Г. Демидов. Без бирки (1966)]

(74) *Сам жил **абы как**, в избенке, не хотел жилое ставить, покуда не разбогатею.* [Сергей Залыгин. Соленая Падь (1967)]

В газетном подкорпусе НКРЯ, отражающем современное словоупотребление, из 333 вхождений единицы *абы* 293 составляют приместоименные обороты.

Заметим, что первоначально *абы* в сочетании с ксловами использовалось преимущественно авторами, родившимися и выросшими в зоне

взаимодействия с украинским или белорусским языком, в которых *абы* активно функционирует как в союзном, так и в приместоименном значении, но малопомалу вошло в общий обиход (хотя, как кажется, до сих пор используется не всеми носителями языка). Современный переводчик не обинуясь вставляет *абы как* в свой перевод произведения Набокова, тогда как сам Набоков едва ли мог употребить такой оборот:

(75) *…the sunset had gone, leaving only a clutter of the purplish remnants of the day, piled up* ***anyhow****, ruins, junk.* [Vladimir Nabokov. Bend Sinister (1947)]
…закат ушел, побросав в беспорядке багровые останки дня, сваленные ***абы как*** *— развалины, хлам.* [Владимир Набоков. Под знаком незаконнорожденных (С. Ильин, 1993)]

4. Выводы

Итак, на примере слова *абы* можно убедиться, что корпусное исследование позволяет не только верифицировать имевшиеся заранее гипотезы и описания, полученные в «докорпусную» эпоху, но и получить совсем новые и часто неожиданные сведения. Когда речь идет о тонких и постепенных семантических изменениях, ценность корпусного материала состоит не столько в получении статистических данных, сколько в возможности наблюдения над медленным изменением словоупотребления и семантики языковых единиц. Так, частица *абы* (в сочетании с *к*-словом) интуитивно кажется старой и даже архаичной. Корпусное исследование показывает, что она совсем новая и что это «архаичное» употребление, предполагающее в современном языке отрицательную полярность, сформировалось в последние несколько десятков лет.

Тем самым тот факт, что, как мы уже отмечали, в «Малом академическом словаре» издания 1981 для слова *абы* указано лишь союзное употребление (и лишь за знаком ромба в качестве фраземы приводится выражение *абы как*), не должен вызывать удивление. Дело не в упущении авторов словаря, а в том, что они ориентировались на картотеку, в основном отражающую словоупотребление XIX и первой половины XX в., когда интересующее нас употребление *абы* еще не стало общепринятым. Однако то, что эта словарная статья практически без изменений перенесена в новое издание «Малого академического словаря» (2016), хотя новое употребление *абы* вошло в обиход уже более полувека тому назад, можно объяснить лишь инерцией восприятия.

Одновременно выяснилось, что высказанная в предшествующих публикациях гипотеза об отрицательной полярности частицы *абы* (в сочетании с *к*словом) подтверждается лишь частично. Оборот «*абы к*-слово» действительно тяготеет к употреблению в контексте эксплицитного или имплицитного отрицания, но эта тенденция вовсе не абсолютна.

Литература

1. *Евгеньева А. П.* (ред.) (1981). Словарь русского языка. Т. 1. М.: Русский язык, 1981.
2. *Крысин Л. П.* (ред.) (2016). Академический толковый словарь русского языка. Т. 1. М.: Издательский дом ЯСК, 2016.
3. *Левонтина И. Б., Шмелев А. Д.* (2005). Малоизученные единицы со значением незаданности критериев выбора в русском языке // Логический анализ языка. Квантификативный аспект языка. М.: «Индрик», 2005. С. 638–651.
4. *Тестелец Я. Г., Былинина Е. Г.* (2005). О некоторых конструкциях со значением неопределенных местоимений в русском языке: амальгамы и квазирелятивы / ИППИ РАН. Семинар «Теоретическая семантика». 15.04.2005 // http://www.rsuh.ru/binary/1787534_99.1322270635.82662.pdf

References

1. *Evgen’eva A. P.* (ed.) (1981). Dictionary of Russian [Slovar’ russkogo yazyka], vol. 1, Russkii yazyk, Moscow.
2. *Krysin L. P.* (ed.) (2016), Academic explanatory dictionary of Russian [Akademicheskii tolkovyi slovar’ russkogo yazyka], vol. 1, Izdatel’skii dom YaSK, Moscow.
3. *Levontina I. B., Shmelev A. D.* (2005), Poorly explored items with the meaning of ‘not specified choice criteria’ in Russian [Maloizuchennye edinitsy so znacheniem nezadannosti kriteriev vybora v russkom yazyke], Logical analysis of language: quantitative aspect of language [Logicheskii analiz yazyka. Kvantifikativnyi aspekt yazyka], Indrik, Moscow, p. 638–651.
4. *Testelets Ya. G., Bylinina E. G.* (2005), On some indefinite pronouns constructions: amalgams and free relatives [O nekotorykh konstruktsiyakh so znacheniem neopredelennykh mestoimenii v russkom yazyke: amal’gamy i kvazirelyativy], paper presented to the workshop “Theoretical semantics”, Moscow, 15.04.2005, http://www.rsuh.ru/binary/1787534_99.1322270635.82662.pdf

ОПЫТ ОБЪЕКТИВНОЙ ОЦЕНКИ ИНТОНАЦИОННОГО КАЧЕСТВА СИНТЕЗИРОВАННОЙ РУССКОЙ РЕЧИ

Лобанов Б. М. (Lobanov@newman.bas-net.by),
Соломенник А. И. (anna.i.prodan@gmail.com),
Житко В. А. (zhitko.vladimir@gmail.com)

Объединённый институт проблем информатики НАН Беларуси, Минск, Беларусь

AN EXPERIENCE OF THE OBJECTIVE ESTIMATION OF INTONATION QUALITY OF THE SYNTHESIZED RUSSIAN SPEECH

Lobanov B. M. (Lobanov@newman.bas-net.by),
Solomennik A. I. (anna.i.prodan@gmail.com),
Zhitko V. A. (zhitko.vladimir@gmail.com)

United Institute of Informatics Problems NAS Belarus, Minsk, Belarus

The paper describes an experiment on an instrumental evaluation of the intonation quality of synthesized Russian speech by using of “Inton@Trainer” computer system. The system was originally designed to train learners in producing the basic intonation patterns of Russian speech. It is based on comparing the melodic portraits of a reference sentence and a sentence pronounced by the learner. Our approach to assessing the intonational quality of speech allows to treat a synthesized speech with the same strict requirements as are applied to students studying Russian as a second language. We describe the technology used for the instrumental evaluation of the intonation quality of synthesized speech and the acoustic database of reference phrases used to assess the intonation quality of synthesized speech. The paper presents the results of testing the intonation quality of two Russian synthetic voices. We discuss the results of the experiment and outline the ways for improving the methods for objective evaluation of synthesized speech prosodic quality, as well as the possibility of applying the developed system in other linguistic tasks.

Keywords: Speech intonation, speech synthesis, objective evaluation, intonation quality, systems of analysis and assessment of intonation, learning system, Russian intonation

Введение

В работе [Lobanov, 2017] дано описание методики использования разработанной ранее компьютерной системы *"Inton@Trainer"* в качестве тренажёра на начальном этапе обучения РКИ в рамках освоения учащимися интонационных конструкций русской речи [Bryzgunova, 1968], [Odintsova, 2011]. Данная работа посвящена описанию эксперимента по нетрадиционному использованию разработанной компьютерной системы *"Inton@Trainer"*[1], а именно в задаче объективной (инструментальной) оценки интонационного качества синтезированной русской речи.

Как известно, существуют две основных разновидности методов оценки качества синтезированной речи: субъективный и объективный (инструментальный). Субъективный метод основываются на статистической обработке субъективных оценок качества синтезированной речи достаточно большим числом слушателей-экспертов.

Для русского языка настоящему времени достаточно разработанным методом оценки интонационного качества синтезированной речи является именно субъективный метод. В работе [Solomennik, 2013] описана методика и результаты субъективного тестирования интонационного качества синтезированной речи отечественных и зарубежных русскоязычных синтезаторов. Для прослушивания аудиторам были предложены 40 тестовых фраз, которые были синтезированы несколькими русскоязычными голосами разных синтезаторов.

В результате общая естественность интонации синтезированных голосов, участвовавших в тестировании, нормированная на естественный голос, изменяется от 49% для голоса «Olga» (Loquendo TTS) до 72% «Владимир» (VitalVoice TTS). Показатели, подсчитанные отдельно для каждого интонационного типа, позволили также найти слабые места каждой системы синтеза в отношении определённых интонационных конструкций. Самый низкий показатель точности был получен для вопросов с вопросительными словами и восклицаний. Длительный опыт работы авторов с синтезированной речью и проведённый эксперимент по влиянию различных ошибок на восприятие слушающими качества синтезированной речи [Solomennik, 2015] показывает, что неадекватное интонирование является основной проблемой качества современных русскоязычных синтезаторов.

Проведение подобных тестов субъективными методами является довольно трудоёмкой задачей, и для того, чтобы ускорить процесс оценки, используются различные инструментальные методы, основывающиеся на автоматическом сравнении синтезированной и естественной речи [Falk, 2008]. Известны также исследования по оценке качества синтезированной речи с использованием систем автоматического распознавания речи [Bachan, 2012]. К объективным методам оценки можно отнести и систему [Norrenbrock, 2012], позволяющую инструментально оценивать качество просодической обработки.

Объективные методы основываются на оценке меры сходства интонации синтезированной и естественной речи с использованием специально разработанных инструментальных средств. В данной работе в качестве такого средства

[1] Больше информации о системе и бесплатное скачивание её демонстрационных версий доступно на сайте: https://intontrainer.by/

предлагается использовать систему *“Inton@Trainer”*, а в качестве эталонного тестового материала использовать ту же фразовую БД [Lobanov, 2017], которая была рекомендована И. В. Одинцовой [Odintsova, 2011] в качестве образцов для освоения учащимися интонационных конструкций русской речи.

Такой подход к оценке интонационного качества речи позволяет отнестись к синтезированной речи со столь же строгими требованиями, как и к учащимся на курсах РКИ.

1. Метод объективной оценки интонационного качества синтезированной речи

В основу методологии оценки интонационного качества положена количественная оценка различных мер сходства для интонационных параметров синтезированной и естественной речи, вычисляемая с помощью системы *“Inton@Trainer”*. Сходство определяется в рамках предложенной ранее [Lobanov, 2014)] модели представления интонационных конструкций русской речи (ИК-1 — ИК-7) — в виде Универсальных (Унифицированных) Мелодических Портретов (УМП) в нормированных координатах «Частота — Время». В качестве критериев сходства интонации синтезированной и естественной речи выступают степень их близости по диапазону изменения частоты основного тона (F0) и форме кривой, отображаемой в виде УМП. Вычисление этих критериев сходста, осуществляемое системой *“Inton@Trainer”*, наглядно отображается на экране. Пример экранного отображения результатов сравнения эталонной интонации (красная линия) с синтезированной (коричневая линия) для фразы «*Она уезжаент за+втра*.» представлен на Рис. 1.

На рисунке красный столбец слева отображает диапазон изменения F0 эталонной естественной фразы, а коричневый — синтезированной. Справа красным цветом отображается линейный график УМП эталонной естественной фразы, а коричневым — синтезированной фразы. Внизу под графиками приведены минимальное и максимальное значения F0 эталонной (Template) и синтезированной (User) фраз. Оценка интонационного сходства синтезированной и естественной речи осуществляется по двум основным критериям: а) по степени сходства диапазонов изменения F0 и б) по степени сходства УМП кривых.

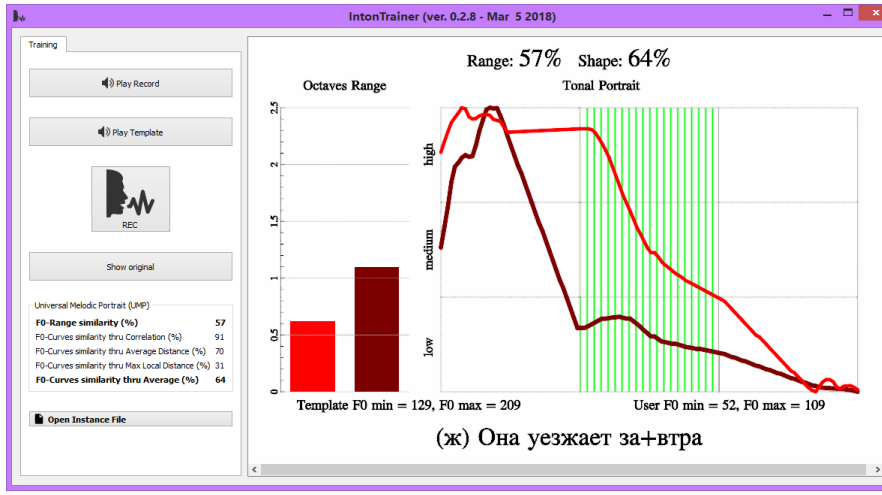


Рис. 1. Результат сравнения интонации фразы «*Она уезжаент за+втра.*» эталонной и синтезированной речи (Голос: Юрий — Nuance Vocalizer)

Вверху над графиками Рис. 1 приведены значения для мер сходства по диапазону — **Range 57%** и по форме кривой мелодического портрета — **Shape 64%**, вычисленные следующим образом.

***А) Оценка сходства диапазонов изменения F0* (Range Similarity).**
Определяется путём вычисления процентных соотношений — F0max / F0min между эталонным (templ) и синтезированным образцом (user) по формуле:

Range Similarity = {(F0max / F0min) user * 100%} / {(F0max / F0min) templ} (1)

***Б) Оценки сходства формы кривых УМП* (Shape Similarity).**
Вычисление **Shape Similarity** осуществляется следующими тремя различными способами.

- ***Корреляционная оценка сходства УМП***. Определяется путём расчёта взаимной корреляции **— r** (см.: http://statistica.ru/) между УМП эталонной и синтезированной фраз. В этом случае мера сходства — **Shape Similarity** определяется по формуле:

  Shape Similarity (thru correlation) = [(r+1) / 2] *100% (2)

- ***Среднее значение сходства УМП***. Определяется путём расчёта среднего значения векторного расстояния **d** между УМП эталонной и синтезированной фраз — Average Distance,

  Shape Similarity (thru average distance) = (1 — D)*100% (3)

- ***Минимальное значение сходства УМП***. Определяется путём расчёта максимального значения локального расстояния — Maximum Local Distance.

  Shape Similarity (thru max local distance) = (1 — dmax)*100% (4)

Использование 3-х различных мер сходства УМП позволяет оценить различие кривых не только в статистическом плане, но также в среднем и на тех участках, где имеется наибольшее отклонение от эталонной кривой.

Таким образом, оценка интонационного качества синтезированной речи, в сравнении с естественной, осуществляется по следующим четырём критериям:

1. Оценка сходства диапазонов изменения F0
2. Корреляционная оценка сходства УМП
3. Среднее значение сходства УМП
4. Минимальное значение сходства УМП

Подробные сведения, касающиеся технологии оценки интонационного качества, приведены на сайте https://intontrainer.by/. (См.: User Guide — Russian: «Анализатор и тренажёр речевой интонации»).

2. Результаты тестирования интонационного качества синтезированной речи

В качестве акустической базы данных эталонных ИК для сравнения с синтезированными фразами были использованы образцы фраз, произносимых диктором с женским голосом, содержащихся в мультимедийном учебнике РКИ [Odintsova, 2011]. Были отобраны 14 простых фраз по две на каждую ИК (см. Табл. 1) таким образом, чтобы по возможности исключить различные варианты прочтения (положение ядра и тип ИК).

Таблица 1. Тексты фраз

ИК1.	*Он гуляет.*	*Она уезжает завтра.*
ИК2.	*Какой?*	*Почему ты опоздал?*
ИК3.	*Любит?*	*Он завтра уезжает?*
ИК4.	*А дедушка?*	*А сегодня?*
ИК5.	*Кому она только не писала!*	*Какой сегодня день!*
ИК6.	*Весело как!*	*Какой фильм!*
ИК7.	*Да какая она актриса!*	*Да какая там выставка!*

Для тестирования были выбраны два наиболее качественных русскоязычных синтезаторов речи с синтезированными мужскими голосами:

- Владимир — VitalVoice TTS (ЦРТ, Россия).
- Юрий — Nuance Vocalize (Nuance Communications, Inc., США).
- Женский голос (Естественная речь).

Женский голос использовался для сравнения интонационного качества синтезированной и естественной речи. Фразы с ИК1 — ИК7 были начитаны диктором-женщиной без предварительного указания требуемой типа ИК.

Для всех вариантов тестирования в качестве эталонных использовались образцы фраз, произносимых диктором с женским голосом, взятые из мультимедийного учебника РКИ.

В таблицах 2, 3 приведены численные значения оценок интонационного сходства синтезированной речи для каждой из 7 интонационных конструкций по четырём вышеописанным критериям, а в таблице 4 — для женского голоса естественной речи.

Таблица 2. Оценки интонационного сходства для синтезатора — VitalVoice

Голос: Владимир				
Тип ИК	**Диапазонная (%)**	**Корреляционная (%)**	**Средняя (%)**	**Минимальная (%)**
ИК-1	22,5	68,5	63,5	32,5
ИК-2	79,5	66,0	65,0	31,5
ИК-3	30,5	78,0	66,0	32,0
ИК-4	36,5	45,0	54,5	19,5
ИК-5	47,5	78,5	72,0	31,0
ИК-6	43,0	53,5	57,5	36,5
ИК-7	23,5	59,5	57,5	22,5
Ср. знач.	40,4	64,1	62,3	29,4

Таблица 3. Оценки интонационного сходства для синтезатора — Nuance Vocalizer

Голос: Юрий				
Тип ИК	**Диапазонная (%)**	**Корреляционная (%)**	**Средняя (%)**	**Минимальная (%)**
ИК-1	35,5	89,0	75,5	34,0
ИК-2	68,0	82,0	73,5	50,0
ИК-3	31,5	90,5	83,5	60,5
ИК-4	54,5	30,0	47,0	13,5
ИК-5	45,5	90,5	78,0	32,0
ИК-6	45,0	50,5	55,0	14,0
ИК-7	13,5	58,0	58,0	33,0
Ср. знач.	41,9	70,1	67,2	33,9

Таблица 4. Оценки интонационного сходства для естественной речи

Женский голос				
Тип ИК	**Диапазонная (%)**	**Корреляционное (%)**	**Средняя (%)**	**Минимальная (%)**
ИК-1	73,0	96,0	79,0	52,5
ИК-2	79,0	82,5	72,5	41,5
ИК-3	22,0	88,5	73,5	39,5
ИК-4	53,5	76,0	69,0	20,0
ИК-5	53,5	95,0	83,5	58,0

Женский голос				
Тип ИК	Диапазонная (%)	Корреляционное (%)	Средняя (%)	Минимальная (%)
ИК-6	61,5	93,0	80,0	49,0
ИК-7	22,0	62,5	56,5	13,0
Ср. знач.	52,1	84,8	73,4	39,1

На рисунке 2 представлены графики усреднённых по ИК1 — ИК7 значений мер сходства с использованием каждого из 4-х видов оценки сходства: 1 — Диапазонной, 2 — Корреляционной, 3- Средней, 4 — Минимальной. Как видно из рисунка, каждый из видов оценки отображает одну и ту же тенденцию увеличения степени сходства в последовательности: Владимир (VitalVoice TTS); Юрий (Nuance Vocalizer); Женский голос (Естественная речь). В данной ситуации трудно отдать предпочтение какому-либо одному из видов оценки сходства. Это решение может быть сделано только в результате более масштабных исследований.

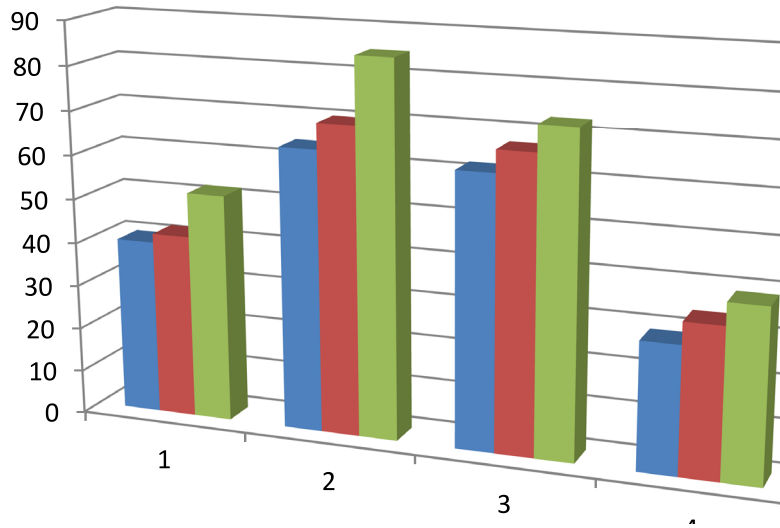


Рис. 2. Усреднённые по ИК1 — ИК7 значения мер сходства для каждого из 4-х видов оценок (Ряд 1 — VitalVoice TTS; Ряд 2 — Nuance Vocalizer; Ряд 3 — Ест. речь)

На рисунке 3 представлены графики усреднённых значений 4-х оценок сходства для каждой из ИК. Как видно из рисунка, тенденция, замеченная при рассмотрении Рис. 2 в основном сохраняется в большинстве случаев. Исключением являются ИК4 и ИК6, когда синтезатор «VitalVoice TTS» показывает лучшие результаты, чем «Nuance Vocalizer». Из рисунка видно также, что преимущество интонационного качества естественной речи, в значительной степени, проявляются в ИК4, ИК6, ИК7 и, в меньшей степени, в ИК2 и ИК3.

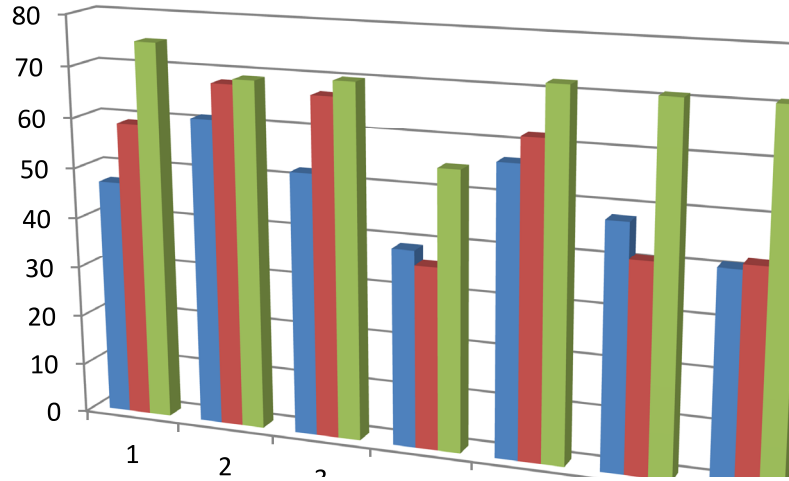


Рис. 3. Усреднённые значения 4-х оценок мер сходства для каждой из ИК (Ряд 1 — VitalVoice TTS; Ряд 2 — Nuance Vocalizer; Ряд 3 — Ест. речь)

Описанные результаты тестирования относятся к простейшим односинтагменным, одноакцентным фразам, которые в реальных текстах представлены лишь незначительным процентом. Практический интерес, однако, в наибольшей степени представляют сложные многосинтагменные фразы, в которых каждая из синтагм может иметь более одного акцента. Именно такого рода фразы повсеместно встречаются в деловых и художественных текстах, лежащих в основе создания звукового сопровождения ТВ-передач и синтезированных аудио книг.

Ниже будут представлены результаты сравнительной оценки интонационного качества естественной и синтезированной речи на примере сложной, многосинтагменной фразы: «Не успела за Андреем затвориться дверь, как я увидел в своем кабинете высокого, широкоплечего мужчину, державшего в одной руке бумажный сверток, а в другой — фуражку с кокардой». В качестве эталона используется соответствующий отрезок сигнала из аудиокниги по повести А. П. Чехова «Драма на охоте» в исполнении профессионального диктора Александра Балакирева. Аудирование этого отрезка показало, что он состоит из 5-ти 3-х акцентных синтагм. Для тестирования были выбраны те же русскоязычные синтезаторы речи с синтезированными женскими и мужскими голосами и естественная речь непрофессионального диктора:

- «Анна» и «Владимир» — VitalVoice TTS (ЦРТ, Россия).
- «Катя» и «Юрий» — Nuance Vocalize (Nuance Communications, Inc., США).
- «Борис» (Естественная речь непрофессионального диктора).

Усреднённые значения оценок интонационных мер сходства с эталонами для каждой из синтагм представлены в таблице 5.

Таблица 5. Усреднённые значения оценок интонационных мер сходства

Текст 5-ти синтагменной фразы: *«Не успела за Андреем затвориться дверь, как я увидел в своем кабинете высокого, широкоплечего мужчину, державшего в одной руке бумажный сверток, а в другой — фуражку с кокардой»*		**Марка синтезатора**				**Диктор**
		Vital Voice TTS		**Nuance Vocalizer TTS**		
		Анна	**Владимир**	**Катя**	**Юрий**	**Борис**
1	*Не успе+ла за Андре+ем затвориться две+рь,*	32	48	74	56	78
2	*как я уви+дел в свое+м кабине+те*	43	41	59	37	81
3	*высо+кого, широкопле+чего мужчи+ну,*	32	36	66	73	81
4	*держа+вшего в одной руке бума+жный све+рток,*	23	38	35	44	78
5	*а в друго+й — фура+жку с кока+рдой.*	57	21	38	76	89
Среднее значение для 5-ти синтагм		**39**	**37**	**54**	**57**	**81**

На рисунке 4 представлены средние значения оценок для 5-ти синтагм по каждому из видов синтезированной и естественной речи.

На рисунках 5 и 6 представлены графические результаты сравнения эталонной интонации (красная линия) фразы с естественной («Борис») и синтезированной речью («Владимир» — VitalVoice TTS).

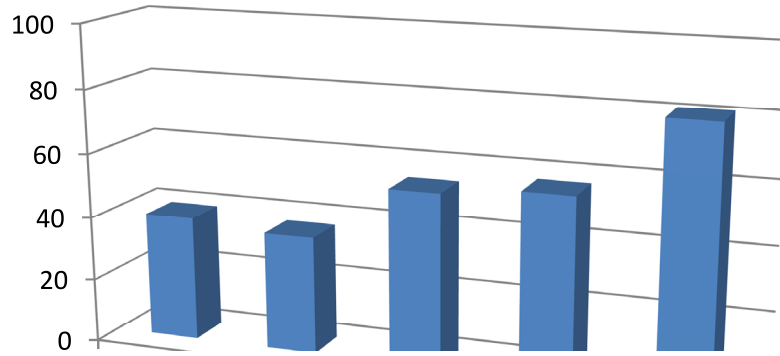


Рис. 4. Средние значения оценок для 5-ти синтагм (Ряд 1, 2 — VitalVoice TTS; Ряд 3, 4 — Nuance Vocalizer; Ряд 5 — Ест. речь)

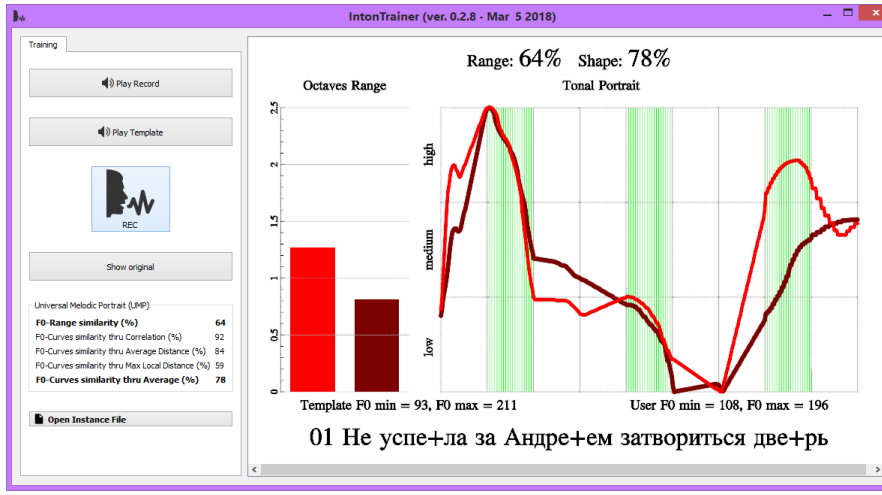


Рис. 5. Результат сравнения эталонной интонации (красная линия) с естественной речью («Борис»)

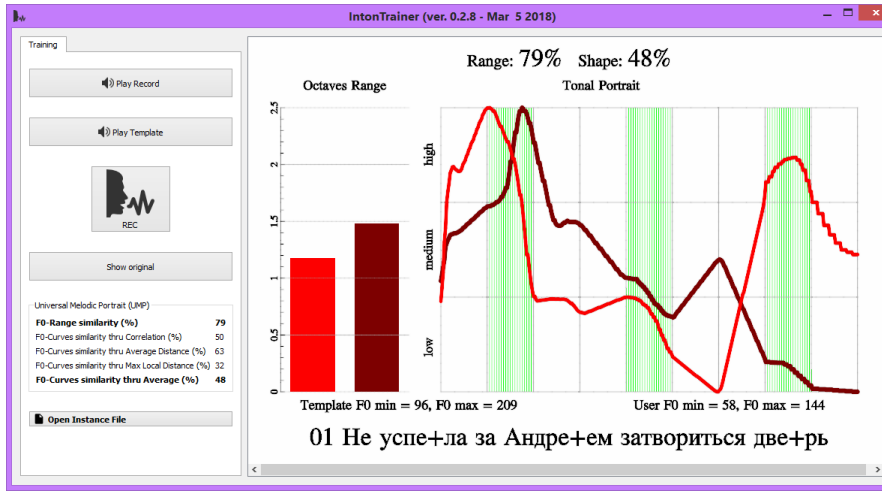


Рис. 6. Результат сравнения эталонной интонации (красная линия) с синтезированной речью («Владимир» — VitalVoice TTS)

Заключение

Представленные в данном докладе результаты объективной оценки интонационного качества синтезированной речи ни в коем случае не следует считать исчерпывающими. Это касается как слишком малого экспериментального материала, так и ограниченного числа исследуемых интонационных параметров. В частности, совершенно отсутствует информация о способах оценки энергетической, ритмической и паузальной компонент просодии речи.

Основным результатом этой работы мы считаем тот факт, что она продемонстрировала возможность использования разработанной системы *“Inton@ Trainer”* не только как нового компьютерного средства при обучении интонации в рамках РКИ, но и как нового компьютерного средства для интонационных исследований.

Литература

1. *Bachan J., Kuczmarski T., Francuzik P.* (2012), Evaluation of synthetic speech using automatic speech recognition, Proc. XIV International PhD Workshop OWD, Wisla, pp. 500–505.
2. *Bryzgunova E. A.* (1968), Sounds and Intonation of Russian Speech [Zvuki I Intonatsiya Russkoy Rechi], Nauka, Moscow.
3. *Falk T. H., Möller S.* (2008), Towards Signal-Based Instrumental Quality Diagnosis for Text-to-Speech Systems, IEEE Signal Processing Letters, Vol. 15, pp. 781–784.
4. *Lobanov B. M., Okrut T.* (2014), Universal Melodic Portraits of Intonation Patterns in Russian Speech [Universalnye melodicheskie portrety intonacionnyh konstrukciy russkoy rechi], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2014” [Komp’yuternaya lingvistika i intellektual’nye tekhnologii: Materialy Mezhdunarodnoy konferentsii “Dialog-2014”], Bekasovo, pp. 330–339.
5. *Lobanov B. M., Zhitko V. A., Kharlamov A. A.* (2017), A computer system of teaching intonation patterns of Russian speech [Komp’yuternaya sistema obucheniya intonatsionnym konstruktsiyam russkoy rechi], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2017” [Komp’yuternaya lingvistika i intellektual’nye tekhnologii: Materialy Mezhdunarodnoy konferentsii «Dialog-2017], Moscow, pp. 287–302.
6. *Norrenbrock C. R., Hinterleitner F., Heute U., Möller S.* (2012), Instrumental Assessment of Prosodic Quality for Text-to-Speech Signals, Signal Processing Letters, IEEE, pp. 255–258.
7. *Odintsova I. V.* (2011), Sounds. Rhythm. Intonation. [Zvuki. Ritm. Intonatsiya.] — Flinta-Nauka, Moscow.
8. *Solomennik A. I., Cherentsova A. E.* (2013), A Method for Auditory Evaluation of Synthesized Speech Intonation, Miloš Železný et al. (Eds.): SPECOM 2013, Lecture Notes in Artificial Intelligence 8113, Springer, pp. 9–16.
9. *Solomennik A. I.* (2015) An influence of defects in synthesized speech on its naturalness [Zavisimost’ estestvennosti zvuchaniya sintezirovannoy rechi ot nalichiya oshibok razlichnykh tipov], Actual problems of philological science: the view of a new generation. Reports of participants of the XX-XXI International conferences of students, graduate students and young scientists «Lomonosov». Section «Philology» [Aktual’nye problemy filologicheskoy nauki: vzglyad novogo pokoleniya. Doklady uchastnikov XX-XXI Mezhdunarodnykh konferentsiy studentov, aspirantov i molodykh uchenykh “Lomonosov”. Sektsiya “Filologiya”], Moscow, pp. 475–480.

EXTRACTING SENTIMENT ATTITUDES FROM ANALYTICAL TEXTS

Loukachevitch N. V. (louk_nat@mail.ru)
Lomonosov Moscow State University, Moscow, Russia

Rusnachenko N. (kolyarus@yandex.ru)
Bauman Moscow State Technical University, Moscow, Russia

In this paper we present the RuSentRel corpus including analytical texts in the sphere of international relations. For each document we annotated sentiments from the author to mentioned named entities, and sentiments of relations between mentioned entities. In the current experiments, we considered the problem of extracting sentiment relations between entities for the whole documents as a three-class machine learning task. We experimented with conventional machine-learning methods (Naive Bayes, SVM, Random Forest).

Keywords: sentiment analysis, coherent texts, relation extraction

ИЗВЛЕЧЕНИЕ ОЦЕНОЧНЫХ ОТНОШЕНИЙ МЕЖДУ СУЩНОСТЯМИ ИЗ АНАЛИТИЧЕСКИХ ТЕКСТОВ

Лукашевич Н. В. (louk_nat@mail.ru)
МГУ имени М. В. Ломоносова, Москва, Россия

Русначенко Н. (kolyarus@yandex.ru)
МГТУ имени Н. Э. Баумана, Москва, Россия

1. Introduction

Automatic sentiment analysis, i.e. identification of opinions on the subject discussed in the text, is one of the most popular applications of natural language processing during last years.

Approaches to extracting the sentiment position from a text depends on the genre of the text being analyzed. So, one of the most studied text genres in the sentiment analysis task is users' reviews about products or services. Such texts are usually

devoted to discussion on a single entity (but, perhaps in its various aspects), and the opinion is expressed by one author, namely the author of the review [Pang et al., 2002; Taboada et al., 2011; Liu 2012; Chetviorkin and Loukachevitch, 2013; Loukachevitch et al., 2015].

Another popular type of texts for sentiment analysis are short messages posted in social networks, especially, in Twitter [Pak, Paroubek, 2010; Loukachevitch, Rubtsova, 2016; Rosenthal et al., 2017]. These texts can require very precise analysis but, at the same time, they cannot contain multiple opinions toward multiple entities because of short length.

One of the most complicated genres of documents for sentiment analysis are analytical articles that analyze a situation in some domain, for example, politics or economy. These texts contain opinions conveyed by different subjects, including the author(s)' attitudes, positions of cited sources, and relations of mentioned entities to each other. Analytical texts usually contain a lot of named entities, and only a few of them are subjects or objects of a sentiment attitude. Besides, an analytical text can have a complicated discourse structure. Statements of opinion can take several sentences, or refer to an entity mentioned several sentences earlier. Also a sentence containing an opinion or describing the relationship's orientation between entities may contain other named entities, which complicates the recognition of sentiment attitudes, their subjects and objects.

In this paper, we present a corpus of analytical articles in Russian annotated with sentiments towards named entities and describe experiments for automatic recognition of sentiments between named entities. This task is a specific subtask of relation extraction.

2. Related Work

The task of extracting sentiments towards aspects of an entity in reviews has been studied in numerous works [Liu 2012, Loukachevitch et al., 2015]. Also extraction of sentiments to targets, stance detection was studied for short texts such as Twitter messages [Amigo et al., 2012; Loukachevitch, Rubtsova, 2016; Mohammad et al., 2017]. But the recognition of sentiments towards named entities or events including opinion holder identification from full texts has been attracted much less attention.

In 2014, the TAC evaluation conference in Knowledge Base Population (KBP) track included so-called sentiment track [Ellis et al., 2014]. The task was to find all cases where a query entity (sentiment holder) holds a positive or negative sentiment about another entity (sentiment target). Thus, this task was formulated as a query-based retrieval of entity-sentiment from relevant documents and focused only on query entities[1].

In [Deng et al., 2015], MPQA 3.0 corpus is described. In the corpus, sentiments towards entities and events are labeled. A system trained on such data should answer such questions as "Toward whom is X negative/positive?", "Who is negative/positive

[1] https://tac.nist.gov/2014/KBP/Sentiment/index.html

toward X"? The annotation is sentence-based. For example, in sentence "When the Imam issued the fatwa against Salman Rushdie for insulting the Prophet...", Imam is negative to Salman Rushdie, Salman Rushdie is negative to Prophet. Imam is also negative toward event of insulting. However, Imam is is positive toward the Prophet. The current MPQA corpus consists of 70 documents. In total, sentiments towards 4,459 targets are labeled.

The paper [Choi, et al., 2016] studied the approach to the recovery of the documents attitudes between subjects mentioned in the text. For example, from the sentence "Russia criticizes Belarus for allowing Mikhail Saakashvili appear on Belarusian TV," it is possible to infer not only the fact that Russia is dissatisfied with Belarus, but also the fact that Russia has the negative attitude toward Mikhail Saakashvili. The authors integrate several attitude constraints into the integer linear programming framework to improve attitude extraction.

To assess the quality of the approach, the text collection consisting of 914 documents was labeled. About 3 thousand opinions were found. As a result of the markup, it was found that about 25% of assessments were extracted not from a single sentence but from neighbor sentences. The best quality of opinion extraction obtained in the work was only about 36% F-measure, which shows that the necessity of improving extraction of attitudes at the document level is significant and this problem is currently studied insufficiently. Inter-annotator agreement was estimated as 0.35 for positive labels and 0.50 for negative labels (Cohen's kappa).

The inference of sentiments with multiple targets in a coherent text, additional features should be accounted for. For the analysis of these phenomena, in the works [Scheible and Schütze, 2013; Ben-Ami et al., 2014] the concept of sentiment relevance is discussed. In [Ben-Ami et al., 2014], the authors consider several types of the thematic importance of entities discussed in the text: the target—the main entity of the text; accidental—an entity only mentioned in this text; relationTarget—the theme of the text is the relation between multiple entities of the same importance; ListTarget—the text discusses several equally important entities sequentially. These types are treated differently in sentiment analysis of coherent texts.

3. Corpus and annotation

In order to initiate research in the field of sentiment analysis of analytical articles for the Russian language, the annotated corpus RuSentRel has been created. The source of the corpus was Internet-portal inosmi.ru, which contains, in the main, analytical articles in the domain of international politics translated into Russian from foreign languages. The collected articles contain both the author's opinion on the subject matter of the article and a large number of sentiments between the participants of the situations described in the article.

For the documents of the assembled corpus, manual annotation of the sentiment attitudes towards mentioned named entities have been carried out. The annotation can be subdivided into two subtypes: 1) the author's relation to mentioned named entities, 2) the relation of subjects expressed as named entities to other named entities. These opinions were recorded as triples: «Subject of opinion, Object of opinion,

sentiment». The sentiment position can be negative (neg) or positive (pos), for example, (*Author, USA, neg*), (*USA, Russia, neg*). Neutral opinions or lack of opinions are not recorded. In contrast to the MPQA 3.0 corpus, the sentiments are annotated for the whole documents, not for each sentence.

In some texts, there were several opinions of the different sentiment orientation of the same subject in relation to the same object. This, in particular, could be due to a comparison of the sentiment orientation of previous relations and current relations (for example, between Russia and Turkey). Or the author of the article could mention his former attitude to some subject and indicate the change of this attitude at the current time. In such cases, it was assumed that the annotator should specify exactly the current state of the relationships.

During the annotation, it became clear that it is very difficult for annotators to indicate all the relationships between named entities, because of the complexity of the texts and the large number of mentioned named entities. Therefore, the procedure of the annotation was as follows: the texts were independently labeled by two annotators, then the annotations were joined, the duplicates were deleted. Duplicates of the attitudes could additionally appear due to different names of the same object (subject), for example, the European Union and the EU. Further, the resulting annotation was checked out by a super-annotator: a small number of discrepancies were resolved, missed relationships could be added. In total, 73 large analytical texts were labeled with about 2000 relations.

To prepare documents for automatic analysis, the texts were processed by the automatic name entity recognizer, based on CRF method [Mozharova, Loukachevitch, 2016]. The program identified named entities that were categorized into four classes: Persons, Organizations, Places and Geopolitical Entities (states and capitals as states). Automatic labeling contains a few errors that have not yet been corrected. Preliminary analysis showed that the F-measure of determining the correct entity boundaries exceeds 95%, there may be some additional errors with the definition of entity types, which is auxiliary information for the sentiment analysis in the current case. In total, 15.5 thousand named entity mentions were found in the documents of the collection.

An analytical document can refer to an entity with several variants of naming (*Vladimir Putin—Putin*), synonyms (*Russia—Russian Federation*), or lemma variants generated from different word forms. Besides, annotators could use only one of possible entity's name in the describing attitudes. For correct inference of attitudes between named entities in the whole document, we provide the list of variant names for the same entity found in our corpus. The current list contains 83 sets of name variants. In such a way, we separate the sentiment analysis task from the task of named entity coreference.

The preliminary version of corpus RuSentRel was granted to the Summer school on Natural Language Processing and Data Analysis[2], organized in Moscow in 2017. The collection was divided into the training and test parts. In the current experiments we use the same division of the data. **Table 1** contains statistics of the training and test parts of the RuSentRel corpus. The last line of the table shows the average number of named entities

[2] https://miem.hse.ru/clschool/

pairs mentioned in the same sentences without indication of any sentiment to each other per a document. This number is much larger than number of positive or negative sentiments in documents, which additionally stress the complexity of the task.

Table 1. Statistics of RuSentRel corpus

	Training collection	Test collection
Number of documents	44	29
Average number of sentences per document	74.5	137
Average number of mentioned entities per document	194	300
Average number of unique named entities per document	33.3	59.9
Average number of positive sentiment pairs of named entities per document	6.23	14.7
Average number of negative sentiment pairs of named entities per document	9.33	15.6
Average number of neutral sentiment pairs of named entities per document	120	276

4. Experiments

In the current experiment we consider the problem of extracting sentiment relations from analytical texts as a three-class supervised machine learning task. All the named entities (NE) mentioned in a document are grouped in pairs: (NE_1, NE_2), (NE_2, NE_1). All the generated pairs should be classified as having positive, negative, or neutral sentiment from the first named entity of the pair (opinion holder) to the second entity of the pair (opinion target). To support this task, we added neutral sentiments for all pairs not mentioned in the annotation and co-occurred in the same sentences into the training and test collections.

As a measure of quality of classification, we take the averaged Precision, Recall and F-measure of positive and negative classes. In the current experiments we classify only those pairs of named entities that co-occur in the same sentence at least once in a document. We use 44 documents as a training collection, and 29 documents as a test collection in the same manner as the data were provided for the Summer School mentioned in the previous section. In the current paper, we describe the application of only conventional machine learning methods: Naive Bayes, Linear SVM and Random Forest implemented in the scikit learn package[3].

The features to classify the relation between two named entities according to an expressed sentiment can be subdivided into two groups. The first group of features characterizes the named entities under consideration. The second group of features describes the contexts in that the pair occurs.

The features of named entities include the following features:

[3] http://scikit-learn.org/stable/

- word2vec similarity between entities. We use the pre-trained model news_2015[4] [Kutuzov, Kuzmenko, 2017]. The size of the window is indicated as 20. Vectors of multiword expressions are calculated as the averaged sum of the component vectors. Using such a feature, we suppose that distributionally similar named entities (for example, from the same country) express their opinion to each other less frequently;
- the named entity type according to NER recognizer: person, organization, location, or geopolitical entity;
- the presence of a named entity in the lists of countries or their capitals. These geographical entities can be more frequent in "expressing sentiments" than other locations;
- the relative frequency of a named entity or the whole synonym group if this group is defined in a text under analysis. It is supposed that frequent named entities can be more active in expressing sentiments or can be an object of an attitude [Ben-Ami et al., 2015];.
- the order of two named entities.

It should be noted that we do not use concrete lemmas of named entities as features to avoid memorizing the relation between specific named entities from the training collection.

The second group of features describes the context in that the pair of named entities is appeared. There can be several sentences in the text where the pair of named entities occurs. Therefore each type of features includes maximal, average and minimum values of all the basic context features:

- the number of sentiment words from RuSentiLex[5] vocabulary: the number of positive words, number of negative words. RuSentiLex contains more than 12 thousand words and expressions with description of their sentiment orientation [Loukachevitch, Levchik, 2016];
- the average sentiment score of the sentence according to RuSentiLex;
- the average sentiment score before the first named entity, between named entities, and after the second named entities according to RuSentiLex;
- the distance between named entities in lemmas;
- the number of other named entities between the target pair;
- number of commas between the named entities.

Altogether, we currently utilize 54 features.

We use several baselines for the test collection: *Baseline_neg*—all pairs of named entities are labeled as negative; *Baseline_pos*—all pairs are labeled as positive, *Baseline_random*—the pairs are labeled randomly; *Baseline_distr*—the pairs are labeled randomly according to the sentiment distribution in the training collection;

[4] http://rusvectores.org/

[5] http://www.labinform.ru/pub/rusentilex/index.htm

Baseline_school—the results obtained by the best team at the Summer school[6]. The results of all baselines are shown in Table 2.

Table 2. Baselines for sentiment extraction between named entities for RuSentRel corpus

Baseline method	Precision	Recall	F-measure
Baseline_neg	0.027	0.390	0.050
Baseline_pos	0.021	0.400	0.040
Baseline_random	0.039	0.215	0.065
Baseline_distr	0.045	0.230	0.075
Baseline_school	0.130	0.103	0.120

Table 3 shows the classification results obtained with the use of several machine learning methods. For two methods (SVM and Random Forest), the grid search of the best combination of parameters was carried out; the grid search is implemented in the same scikit-learn package. The best results were obtained with the Random Forest classifier. The parameter tuning did not improve the results, which are quite low.

Table 3. Results of sentiment extraction between named entities

Method	Precision	Recall	F1
KNN	0.18	0.06	0.09
Naïve Bayes Gauss	0.06	0.15	0.11
Naïve Bayes Bernoulli	0.13	0.21	0.16
SVM (Default values)	0.35	0.15	0.15
SVM (Grid search)	0.09	0.36	015
Random forest (Default values)	0.44	0.19	0.27
Random forest (Grid search)	0.41	0.21	0.27

But we can see that the baseline results are also very low. It should be noted that the authors of the [Choi, et al., 2016], which worked with much smaller documents, reported F-measure 36%.

There is an important question about inter-annotator agreement because of the complexity of the task. In our case, the procedure is not straightforward, because we asked people to indicated only positive or negative relations between named entities but in fact they internally classified the relations into three classes including neutral. Because of the large number of the mentioned named entities in the texts, neutral relations significantly prevail.

We used the following approach. We asked another super-annotator (see Section 3) to label the collection, and compared her annotation with our gold standard using average F-measure of positive and negative classes in the same way as for automatic approaches. In such a way, we can reveal the upper border for automatic

[6] https://miem.hse.ru/clschool/results

algorithms. We obtained that F-measure of human labeling is 0.55. This is quite low value, but it is significantly higher than the results obtained by automatic approaches. About 1% of direct contradictions (positive vs. negative) among all etalon labels were found.

5. Analysis of errors

In this section we consider several examples of erroneous classification of relations between entities.

1) In the following example, the system did not detect that Liuhto is positive towards NATO. This is because of relatively long distance between Liuhto and NATO and absence of evident sentiment words.

   Лиухто говорит, что он начал склоняться к вступлению Финляндии в НАТО.
   (*Liuhto says that he began to welcome Finland's accession to NATO.*)

2) In the following sentence, the evident sentiment words are also absent, and the system misses the sentiment from Putin to Russia,

   Путин хочет войти в историю как царь, расширивший территорию России.
   (*Putin wants to go down in history as the king who expanded the territory of Russia.*)

3) In the following text fragment, there are several sentences about the stance of Sven Mikser towards Putin. But in the first sentence, his position is expressed too complex. The relation is discussed also in the next sentences but pronoun resolution is needed:

   Глава комиссии по иностранным делам эстонского парламента Рийгикогу, бывший министр иностранных дел Свен Миксер (Sven Mikser) считает, что, возможно, президент Владимир Путин не стремится присоединить к России в первую очередь страны Балтии, но подобные намерения вполне могут существовать.

   The head of the Foreign Affairs Committee of the Estonian Parliament, the Riigikogu, the former Foreign Minister Sven Mikser believes that, perhaps, President Vladimir Putin does not seek to join the Baltic countries first of all, but such intentions may well exist.

We can see the large number of long distances between entities having positive or negative attitudes to each other. So, our next step is experiment with convolutional neural networks, which can represent such word sequences using multiple filters. To enhance our training collection semi-automatically, we plan to gather sentences describing known relations, for example, Russia—Ukraine, or United Stated—Bashar Asad. But most such relations are negative.

Conclusion

In this paper we presented the RuSentRel corpus including analytical texts in the sphere of international relations. For each document, we annotated sentiments from the author to mentioned named entities, and sentiments of relations between mentioned entities.

In the current experiments, we considered the problem of extracting sentiment relations between entities for the whole documents as a three-class machine learning task. We experimented with conventional machine-learning methods (Naive Bayes, SVM, Random Forest). The corpus RuSentRel is published[7].

Our next step is experiments with convolutional neural networks, which can represent long word sequences using multiple filters. We plan to enhance our training collection semi-automatically, trying to find sentences describing known relations, to obtain enough data for training neural networks.

Acknowledgments

This work is partially supported by RFBR grant № 16-29-09606.

References

1. *Amigo E., Albornoz J. C., Chugur I., Corujo A., Gonzalo J., Martin T., Meij E., de Rijke M., Spina D.* (2013), Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems, CLEF 2013, Lecture Notes in Computer Science Volume 8138, pp. 333–352.
2. *Ben-Ami Z., Feldman R., Rosenfeld B.* (2014), Entities' Sentiment Relevance, In proceedings of ACL-2014, pp. 87–92.
3. *Ben-Ami Z., Feldman R., Rosenfeld B.* (2015), Exploiting the Focus of the Document for Enhanced Entities' Sentiment Relevance Detection, Data Mining Workshop (ICDMW), 2015 IEEE International Conference on, IEEE, pp. 1284–1293.
4. *Chetviorkin I., Loukachevitch N.* (2013), Evaluating sentiment analysis systems in Russian, In Proceedings of ACL 2013, 12.
5. *Choi E., Rashkin H., Zettlemoyer L., Choi, Y.* (2016), Document-level Sentiment Inference with Social, Faction, and Discourse Context. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL-2016, pp. 333–343.
6. *Deng L., Wiebe J.* (2015), Mpqa 3.0: An entity/event-level sentiment corpus // Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1323–1328.
7. *Ellis J., Getman J., Strassel S. M.* (2014), Overview of linguistic resources for the tac kbp 2014 evaluations: Planning, execution, and results, Proceedings of TAC KBP 2014 Workshop, National Institute of Standards and Technology, pp. 17–18.

[7] https://github.com/nicolay-r/RuSentRel

8. *Kutuzov A., Kuzmenko E.* (2017), WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, vol 661. Springer, Cham.
9. *Liu B.* (2012), Sentiment analysis and opinion mining (synthesis lectures on human language technologies). Morgan & Claypool Publishers.
10. *Loukachevitch N., Rubtsova Y.* (2016), SentiRuEval-2016: Overcoming Time Gap and Data Sparsity in Tweet Sentiment Analysis, Computational Linguistics and Intellectual Technologies Proceedings of the Annual International Conference Dialogue, Moscow, RGGU, pp. 416–427.
11. *Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Y., Ivanov V., Tutubalina E.* (2015), SentiRuEval: Testing Object-oriented Sentiment Analysis Systems in Russian, Proceedings of International Conference of Computational Linguistics and Intellectual Technologies Dialog-2015, V. 2, pp. 2–13.
12. *Loukachevitch N. V., Levchik A.* (2016), Creating a General Russian Sentiment Lexicon, In Proceedings of LREC-2016.
13. *Mohammad S. M., Sobhani P., Kiritchenko S.* (2017), Stance and sentiment in tweets //ACM Transactions on Internet Technology (TOIT).V 17, №. 3, pp. 26.
14. *Mozharova V. A., Loukachevitch N. V.* (2016), Combining knowledge and CRF-based approach to named entity recognition in Russian, International Conference on Analysis of Images, Social Networks and Texts, Springer, Cham, pp. 185–195.
15. *Pak A., Paroubek P.* (2010), Twitter as a Corpus for Sentiment Analysis and Opinion Mining, In proceedings of LREC-2010, pp. 1320–1326.
16. *Pang B., Lee L., Vaithyanathan S.* (2002), Thumbs up?: sentiment classification using machine learning techniques, Proceedings of the ACL-02 conference on Empirical methods in natural language processing, V. 10, pp. 79–86.
17. *Rosenthal S., Farra N., Nakov P.* (2017), SemEval-2017 task 4: Sentiment analysis in Twitter, In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 502–518.
18. *Scheible C., Schütze H.* (2013), Sentiment Relevance. In Proceedings of ACL 2013 (1), pp. 954–963.
19. *Taboada M., Brooke J., Tofiloski M., Voll K., Stede M.* (2011), Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), pp. 267–307.

РЕИНТЕРПРЕТАЦИЯ СОБЫТИЯ: НАБЛЮДЕНИЯ НАД ОДНОЙ РУССКОЙ ЯЗЫКОВОЙ ИННОВАЦИЕЙ[1]

Лютикова Е. А. (lyutikova2008@gmail.com),
Татевосов С. Г. (tatevosov@gmail.com)
МГУ имени М. В. Ломоносова, МПГУ, Москва, Россия

RE-INTERPRETING EVENTS: NOTES ON ONE LINGUISTIC INNOVATION IN RUSSIAN[2]

Lyutikova E. A. (lyutikova2008@gmail.com),
Tatevosov S. G. (tatevosov@gmail.com)
Lomonosov Moscow State University, Moscow Pedagogical State University, Moscow, Russia

The paper explores the distribution and interpretation of the discourse marker *po(-)xodu* (PX) and addresses a possible path of its diachronic development. We argue that the range of uses of PX attested in the corpora supports an analysis that identifies three meanings / functions of this item labeled eventive PX, epistemic PX and discourse-level PX throughout this paper. We propose that the latter two are the products of re-interpretation of the former. We argue for a presuppositional analysis of the eventive PX whereby it requires there be a set of background events that show a temporal overlap with the asserted event and add up to the integral whole. We analyze the epistemic PX as resulting from inferential reinterpretation of the relationship between background and asserted events, with the abductive reasoning being the key ingredient of this reinterpretation. Finally, we treat the discourse-level PX as a counterpart of the eventive PX in the domain of speech acts. We speculate that Krifka's (2014) recent view of speech acts as index changers opens a way of accounting for this parallelism in a principled way. On the diachronic side, we identify PX as the product of diachronic development of the construction in which the argument of the noun *xod* 'move' is expressed by an overt DP. In the course of development,

[1] Данное исследование поддержано грантом РНФ (проект 16-18-02003 «Структура значения и ее отображение в системе лексических и функциональных категорий русского языка», реализуемый в МПГУ).

[2] The research has been supported by Russian Science Foundation (project 16-18-02003 "Structure of meaning and its mapping into lexical and functional categories of Russian" at MPSU).

this DP was first replaced by *pro*, which gave rise to the eventive PX, and later on developed epistemic and discourse-level meanings / functions.

Keywords: event semantics, discourse structure, discourse particles, inferentiality, speech acts

1. Введение

Цель этих заметок — охарактеризовать семантику и дистрибуцию выражения *по ходу / походу / по-ходу* (далее ПХ), получившего массовое распространение настолько недавно, что у М. А. Кронгауза не было возможности включить его в обширный каталог инноваций [Кронгауз 2008]. Характерные примеры употребления этого выражения показаны в (1a–c).

(1) a. *Современная молодежь в поисках своего лица бесконечно придумывает неологизмы, **по ходу** изобретая диалекты XXI века.* [Много японских языков // «Русский репортер», 2012]

b. [Roberta, жен] *И надо ехать на дачу, закруглять» стройку «, а там по агентурным данным» конь не валялся «… и придёт трындец моему так нелегко добытому спокойствию :* (Короче, деньги кончились, кухня ***походу*** только в следующем году :)*]* # [Форум: Апгрейд дачной кухни (2011–2013)]

c. *Твои предки наверняка помнят популярный в 60-е гг. фильм «Его звали Роберт». **По ходу**, это была первая отечественная экранизация на роботему.* [Top 10 веселых роботов (2004) // «Хулиган», 2004.08.15]

Примеры такого типа создают стандартный для семантического исследования набор вопросов. Каков семантический вклад выражения ПХ в интерпретацию целого предложения и — шире — целого фрагмента дискурса? Сколько значений — или, в терминах Московской семантической школы, лексико-семантических вариантов — следует предположить у ПХ, чтобы объяснить всё наблюдаемое разнообразие семантических эффектов? Если значений более одного, какими принципами или правилами определяется их взаимодействие? Прослеживаются ли закономерности в диахроническом развитии ПХ, и в чем они состоят?

В последующих разделах мы попытаемся наметить предварительные обобщения по каждой из этих позиций. Вот краткий конспект разделов 2–3. ПХ имеет пресуппозициональный характер. Есть основания выделить у ПХ три значения, условно называемые событийным, эпистемическим и дискурсивным. Дискурсивное значение возникает из событийного посредством переноса одной и той же пресуппозиции из области событий в область пропозиций. Эпистемическое значение получается посредством инферентивной реинтерпретации событийного значения. Событийное значение диахронически первично и фиксируется в корпусах с конца 19 века. Два других значения представляют собой инновацию начала 21 века.

2. Эмпирические обобщения

В этом разделе мы представим содержательные обобщения, касающиеся дистрибуции и значений ПХ. Эти обобщения станут основой теоретико-модельного анализа семантики ПХ в разделе 3.

2.1. Три значения

Основной тезис, которым определяется последующий анализ ПХ, сформулирован в (2).

(2) *Употребления ПХ без остатка распадаются на три группы: событийное, эпистемическое и дискурсивное.*

(1a) иллюстрирует событийное ПХ. Говоря неформально, при таком употреблении предложение сообщает, что содержащееся в ассерции событие, или утверждаемое событие, происходит параллельно некоторому другому событию или событиям, которые мы далее называем фоновыми. Фоновые события могут быть упомянуты в предшествующем дискурсе или выводиться из контекста.

В (1a) в сфере действия ПХ находится событие, удовлетворяющее дескрипции ‘молодежь изобретает диалекты XXI века’. Фоновым событием выступает ‘молодежь бесконечно придумывает неологизмы’. Для тех и других событий принципиально важно временно́е пересечение (в предельном случае — совпадение во времени): утверждаемое событие должно иметь место «по ходу» фонового. В (3) дескрипции утверждаемого и фонового событий помечены, соответственно, полужирным и подчеркиванием.

(3) a. [ashtu, nick] *Сама перечитывала у Спока про детские болезни на той неделе и* ***прочла*** *по-ходу* ***еще несколько глав****.* [Форум: Доктор Спок (2012)]

b. *Приятель мой думу думает, а по ходу* ***ведёт допрос с пристрастием на предмет анамнеза!*** [Татьяна Соломатина. Акушер-ХА! Байки (2009)]

c. *Слухи о ее делах и проделках просачивались сквозь железный занавес,* ***обрастая*** *по ходу* ***фантастическими подробностями****.* [Вячеслав Борисов. Это мама. Нина Хаген взорвет бункер (2002) // «Известия», 2002.10.13]

Другой важный, хотя и трудно формализуемый семантический компонент событийного ПХ можно обозначить как событийная когерентность. Утверждаемое и фоновое события должны объединяться в естественное целое или образовывать естественный класс. Рассмотрим пример (4). В (4) в сферу действия ПХ входит событие ‘собачья еда и многие другие предметы тоже летают по комнате’. Фоновое событие из предшествующего контекста — ‘мы с мужем молча его раздеваем, одеваем в уличный комбинезон на голое тело, все остальное вешаем на батарею’. Вся последовательность в целом составляется в описание нервного эпизода семейной жизни, существенными компонентами которого выступают и фоновое, и утверждаемое событие.

(4) *Мы с мужем молча его раздеваем,* [все] *одеваем в уличный комбинезон на голое тело, все остальное вешаем на батарею.* ***По ходу*** *собачья еда и многие другие предметы тоже летают по комнате.* [Форум: Родственники мужа разрушают границы ребенка (2013)]

Еще две иллюстрации этого же эффекта показаны в (5a–b). В (5a) фоновое и утверждаемое события сополагаются как частные проявления неприятных автору черт русского характера. В (5b) находка снаряда времен войны подается как событие, естественным образом сопровождающее контртеррористическую операцию (хотя, конечно, не являющееся ее непременным компонентом).

(5) a. *Пустите нас еще раз в Европу, и мы продолжим приготовление щавелевого супа в биде.* ***По ходу*** *мы и фирмачей в два счета принудим справлять нужду в лифте.* [Иван Охлобыстин. Мозг Ихтиандра (1997) // «Столица», 1997.04.15]

b. *В ходе принятых мер заложники были освобождены, террористы нейтрализованы. Уничтожены и заложенные ими взрывные устройства. А заодно и найденный* ***по ходу*** *снаряд времен войны.* [Никита Юрьевский. Тяжело в ученье (2013.05.17) // «Новгородские ведомости», 2013]

Второй тип употребления ПХ иллюстрируется в (1b), повторяемом здесь как (6).

(6) [Roberta, жен] *И надо ехать на дачу, закруглять» стройку «, а там по агентурным данным» конь не валялся «… и придёт трындец моему так нелегко добытому спокойствию :* (Короче, деньги кончились, кухня ***походу*** только в следующем году :)*] #* [Форум: Апгрейд дачной кухни (2011–2013)]

Для такого употребления противопоставление утверждаемого и фонового событий по-прежнему остается релевантным: в (6) это, соответственно, ‘покупка кухни откладывается на следующий год’ и ‘деньги кончились’. Однако отношение между этими событиями в (6) отличается от примеров типа (3)–(5) по двум важнейшим параметрам. Во-первых, одновременности событий более не требуется; во-вторых, фоновое событие выступает эпистемическим основанием для утверждаемого. В (6) время осуществления события ‘деньги кончились’ (прошлое) очевидным образом не пересекается со временем покупки кухни (будущее). А утверждение о том, что покупка кухни произойдет в следующем году, опирается на логический вывод, ключевым компонентом которого выступает пропозиция ‘деньги кончились’. Еще несколько иллюстраций представлены в (7a–d).

(7) *a. Салон не потерся, пластик в салоне не гремит и не скрипит* (мечта владельцев отечественных ТАЗиков:)*), шумоизоляция неплохая, кондиционер работает.* ***По ходу****, где-то сифонит глушак, шума стало многовато.* [Внедорожник для хулигана: Terrano II (2004) // «Хулиган», 2004.07.15]

b. [hoseraul, nick] *Подмосковье — самый богатый регион в заМКАДье. Уж поверьте мне!* [designermars, nick] *Ты в других регионах не был* ***походу****.* [Форум: Разорять область они не боялись, а книжку боятся (2011)]

c. [Ксения, nick] *Походу девочке нет и 16? И о чем тогда разговор? С какой целью сюда пришла?* [Ianika, nick] *С целью, что умные и знающие люди написали, что знают о таких делах… а ты* ***походу*** *не особо одаренная, если что!!!! чтоб быть модератором, надо быть как минимум образованным и правильно излагать свои мысли…!!!!!!* [Форум: Помогите пожалуйста! Расскажите про СИЗО №4. Москва! (2012)]

d. *Слушай, а в правом <ботинке> еще вода. Ты,* ***по ходу****, не все вылил.* [Андрей Геласимов. Год обмана (2003)]

Примечательным образом, все эти примеры допускают синонимичные аналоги с выражениями, вводящими эпистемическую модальность или эвиденциальность — *видимо*, (*как*) *я вижу*, *оказывается*, *судя по всему, похоже*[3]:

(8) a. *Видимо, где-то сифонит глушак, шума стало многовато.*
b. *Ты в других регионах не был,* (как) *я вижу.*
c. *А ты, судя по всему, не особо одаренная!*
d. *Ты, оказывается, не все вылил.*

Все такие выражения имеют в своей семантике инференциальный компонент. Хотя для каждого из них детали того, как он конструируется, могут быть различны, общая идея инференциальности достаточно прозрачна: фоновые пропозиции подаются как необходимое условие для ассерции. В (7a)-(8a) истинность пропозиции ‘шума стало многовато’ выступает основанием для утверждения ‘где-то сифонит глушак’. В (7b)-(8b) фоновым событием выступает речевой акт собеседника (Подмосковье — самый богатый регион заМКАДья); наличие этого события становится основанием для говорящего утверждать ‘ты в других регионах не был’. Аналогично в (7c)-(8c) доступное в контексте (речевое) поведение собеседника обосновывает утверждение ‘ты не особо одаренная’. Наконец, в (7d)-(8d) утверждение о наблюдаемом положении дел (в правом ботинке вода) указывает на истинность пропозиции ‘ты не все вылил’. Более эксплицитную характеристику эпистемической интерпретации ПХ мы дадим в разделе 3.2[4].

---

[3] Многие участники интернет-обсуждений ПХ (а также два из трех рецензентов «Диалога») высказывают предположение о происхождении эпистемического ПХ от наречия похоже путем паронимической аттракции. Не исключая возможного влияния фонетического сходства на развитие эпистемического значения у ПХ, мы, тем не менее, предложим анализ, в соответствии с которым это значение предсказуемым образом развивается из событийного. При таком анализе эпистемическое значение возникает не из-за поверхностного сходства двух лексических единиц, а как результат работы хорошо известного механизма семантической деривации.

[4] Рецензент «Диалога» указывает на желательность сопоставить ПХ «с употреблениями других дискурсивных слов, которые могут выступать в качестве их квазисинонимов». Мы полностью согласны с этой идеей. Мы рассматриваем ее, однако, как задачу на будущее

Третий тип интерпретации, который мы предполагаем у ПХ, иллюстрируется в (1c), повторяемом как (9a), а также в (9b-d).

(9) a. *Твои предки наверняка помнят популярный в 60-е гг. фильм «Его звали Роберт». **По ходу**, это была первая отечественная экранизация на роботему.* [Top 10 веселых роботов (2004) // «Хулиган», 2004.08.15]

b. *Но если по делу, то так и не последовало* (ни от кого) *объяснений, чем именно качество изображения у современного ЖК Самсунга, хуже ЖК Панасоник или Филипс? Я не фанат Самсунга, но все ж смешно слышать голословные заявления, основанные, наверное, на собственном снобизме! Да и **по ходу**, может Джива напомнит, ЖК матрицы какого производителя используют например Панасоник или Сони в своих ТВ? Думаю, не открою секрет, что качество изображения у ЖК ТВ … в первую очередь зависит от МАТРИЦЫ….* [http://corpus.leeds.ac.uk]

c. *Правда, меня с детства, как только прочитал книжку про Миклухо-Маклая, интересовало, почему папуасов не называют неграми* (а чукчей, ***по ходу***, не называют азиатами, а только и исключительно — чукчами). [http://www.trworkshop.net]

d. *Исполнитель: Fun Lovin' Criminals. Кайфоухость: 9 звезд. Понтовые Нью-Йоркские пацаны, некисло задвигающие по тамошним клубам, свое дело понимают туго. NY, **по ходу**, один из самых многонациональных городов в и без того разноцветных Штатах. Так что на вечеринах собирается самая разная публика)* [Макар Свирепый. Байда: музыка (2003) // «Хулиган», 2003.12.15]

Этот тип, как представляется, значительно отличается и от событийного ПХ, и от эпистемического ПХ. В (9a–d) с интуитивной точки зрения речь не идет о соотнесении разворачивающихся одновременно событий, как в случае с событийным ПХ. Невозможно и анализировать эти примеры как указывающие на инференциальные отношения между пропозициями. В (9d), например, не предполагается, что пропозиция 'Нью-Йорк — один из самых многонациональных городов' — это результат логического вывода, подкрепляемого фоновой пропозицией или пропозициями. Мы полагаем, что функция ПХ в этом классе примеров содержательно аналогична функции событийного ПХ, однако реализуется не в области событий, а в области речевых актов. ПХ P сообщает, что речевой акт P надлежит рассматривать на фоне уже развернутой последовательности фоновых речевых актов P_1, … P_n. Текущий речевой акт должен образовывать с фоновыми P_1, … P_n естественное целое. Предельно упрощая, дискурсивное ПХ P можно свести к следующему неформальному толкованию:

и не будем пытаться приступить к решению в пределах этой статьи. Составление лексикографического портрета русских эпистемических обстоятельств и их сравнение — слишком большой и многоплановый сюжет, чтобы его можно было охватить на нескольких страницах.

параллельно развертыванию последовательности речевых актов P_1, … P_n и в дополнение к ним я совершаю речевой акт P.

В (9a) имеется, по-видимому, единственный фоновый акт: утверждение ‘твои предки помнят фильм «Его звали Роберт»’. ПХ, вводящее следующее утверждение ‘это была первая отечественная экранизация на роботему’, сигнализирует, что оба утверждения образуют когерентную дискурсивную общность. Аналогичные наблюдения можно сделать и о других предложениях в (9).

2.2. Источники и полисемия

Хотя официальная лингвистика пока не включила ПХ как самостоятельную единицу в словари современного русского языка (соответствующая словарная статья имеется только в [Никитина 2003]), вопрос об употреблении этого выражения, особенно в речи молодежи, волнует носителей языка. Многочисленные интернет-форумы[5] в последние годы обсуждают происхождение и значения ПХ и его правописание — следует ли писать это выражение слитно, раздельно или через дефис, требуется ли выделять запятыми[6]. В Викисловаре словарная статья ПОХОДУ появляется 16 марта 2011 года[7].

Следует, по-видимому, согласиться с интуицией носителей языка: ПХ представляет собой результат опущения генитивного зависимого в выражении ‘по ходу Х-а’. С предлогом *по* существительное *ход* образует коллокации в двух значениях: в значении ‘в направлении движения’ (примеры (10a–b)) и в значении ‘по мере развертывания’ (толкования (11a–b) и примеры (11c–d)).

(10) a. *Конечно, пирамидальные тополя слева* ***по ходу машины*** *сильно вытянулись, а в остальном дорога мало чем изменилась.*
[Вацлав Михальский. Прощеное воскресенье // Октябрь, 2009]

b. *Состав для мытья наливают в чистую мелкую посуду и протирают пластинку* ***по ходу звуковых дорожек*** *рулончиком нового, желательно стерильного бинта.* [Домашние заботы // «Химия и жизнь», 1985]

---

[5] См., например, https://rus.stackexchange.com/questions/25154/Новомодное-словечко-по-ходу-или-походу/420396, http://www.bolshoyvopros.ru/questions/35376-chto-oznachaet-fraza-po-hodu.html, https://pishu-pravilno.livejournal.com/3730613.html, http://www.trworkshop.net/forum/viewtopic.php?f=55&t=40870, https://pikabu.ru/story/kak_pravilno_po_khodu_ili_pokhodu_po_khodu_pridetsya_zapilit_otdelnyim_postom_3917474, https://eva.ru/static/forums/77/2006_1/538485.html.

[6] Любопытно, что вырабатывается стихийная тенденция различения значений ПХ на письме. Во-первых, эпистемическое и дискурсивное ПХ, в отличие от событийного, часто выделяются запятыми, как вводные слова. Во-вторых, для эпистемического и дискурсивного ПХ значительно чаще используется слитное написание (походу) или написание через дефис (по-ходу). По-видимому, в сознании носителей подобная орфографическая норма связана с конверсией предложно-именных конструкций в наречия (ср. по-новому, подолгу).

[7] https://ru.wiktionary.org/wiki/%D0%BF%D0%BE%D1%85%D0%BE%D0%B4%D1%83

(11) a. *ХОД. <…> **3.** ед.* (в, на ходе)*, перен. Развитие, развёртывание чего-н. Х. событий. Х. войны. По ходу дела* (по обстоятельствам дела) [Ожегов, Шведова 2006].

b. *2.* (в, на ходе) *перен., только ед. Развитие, течение чего-нибудь… Ход событий. Ход войны. По ходу дела обнаружились недочеты. При таком ходе дела. Перерыв в ходе переговоров…* [Ушаков 2001].

c. *Я вчера твоему аппарату профилактику делал, так я **по ходу дела**, видимо по пьяни, кой-какие контакты перепутал.* (Интернет-корпус LEEDS [http://corpus.leeds.ac.uk/ruscorpora.html])

d. *<…> он глубоко и систематически изучал историю и философию и делал **по ходу чтения** аналитические записи.* (Интернет-корпус LEEDS [http://corpus.leeds.ac.uk/ruscorpora.html])

Очевидно, что именно второй вариант является источником событийного ПХ. Поиск коллокаций вида ‘по ходу S_{GEN}’ показывает, что наибольший ранг[8] имеют существительные, обозначающие действия или процессы, имеющие определенную временную протяженность: *дело*, *действие*, *игра*, *повествование*, *событие*, *работа*, *разговор*, *чтение*, *изложение*, *сезон*, *сюжет*, *фильм*, *съемки*, *турнир* и т.д. Кроме того, в качестве дополнения существительного *ход* используется также предикативная структура, вложенная в функциональную оболочку указательного местоимения (*по ходу того, как / что…*, ср. (12a-b)).

(12) a. *Скажешь один раз — нужно сказать и другой, и третий, и четвертый, т.е. комментировать **по ходу того, что происходит**.* [Рой Медведев. Долгий путь домой. Солженицын и перестройка (2002)]

b. ***По ходу того как облачают себя в сценические одеяния**, спина как-то выпрямляется, появляется осанка; накладывается умеренный грим…* [Нонна Мордюкова. Казачка (2005)]

По-видимому, ПХ возникает в результате контекстно или ситуативно лицензируемого выражения генитивного зависимого нулевым анафорическим местоимением (*pro*). Первое зафиксированное в НКРЯ употребление событийного ПХ, датируемое концом 19 в., устроено именно так, ср. (13). Анафорическое происхождение имплицитного аргумента выражения ‘по ходу Х-а’ находит отражение в пресуппозициональном характере семантического компонента, привносимого ПХ (см. 2.3).

(13) *Ему очень хорошо удавались роли слуг. А в драме Н. А. Потехина «Быль молодцу не укор» он имел неимоверно большой успех, имитируя известного капельмейстера Иоганна Штрауса, в то время только что прославившегося в Петербурге. **По ходу**, одна из декораций изображала*

[8] Поиск коллокаций осуществлялся на материале интернет-корпуса русского языка в программе подсчета коллокаций Центра по изучению перевода университета г. Лидс (http://corpus.leeds.ac.uk/), в качестве меры использовалась логарифмическая функция правдоподобия (LogLikelihood score).

Павловский вокзал с эстрадой, на которой появлялся Антонолини с оркестром, дирижируя которым он так безукоризненно подражал оригиналу.… [А. А. Нильский. Закулисная хроника. (1856–1894)].

Эпистемическое и дискурсивное ПХ впервые обнаруживаются в НКРЯ только в текстах 21-го века. В Таблице 1 показано количество употреблений ПХ в разных значениях в двух подкорпусах: до 2000 г. включительно и после 2000 г.[9]

Таблица 1. Распределение значений ПХ в НКРЯ

	событийное ПХ	эпистемическое ПХ	дискурсивное ПХ
до 2000 г. включительно	25	0	0
после 2000 г.	57	17	8

Такое соотношение наводит на мысль, что эпистемическое и дискурсивное ПХ являются производными от событийного. Если это так, можно проследить направление и тип семантической деривации, образующей производные значения.

Мы полагаем, что эпистемическое и дискурсивное значения развиваются параллельно и за счет разных семантических операций. Эпистемическое значение возникает вследствие метафорического переосмысления временно́го пересечения утверждаемого и фонового событий как их импликативной связи в контексте событийной когерентности. Дискурсивное значение развивается из метатекстового использования событийного ПХ в контекстах, где говорящий характеризует собственный речевой акт (14). При использовании вне мета-контекста, как в (15), временно́е пересечение и когерентность фонового и утверждаемого событий метонимически переносятся на временно́е пересечение и дискурсивную когерентность фоновых пропозиций и ассерции.

(14) a. *А можно я* ***спрошу по ходу****… Сколько жердочки могут находиться в использовании в клетке?* [http://parrots.ru/threads/Памятка-начинающему-владельцу-попугая.12798/page-2]

b. *Leonid_B, спасибо! Но тогда* ***по ходу еще один вопрос****. После выполнения запроса <…> наблюдаю, что много таблиц, которые не относятся к служебным схемам, в полях last_analyzed и stattype_locked имеют null.* [http://www.sql.ru/forum/734330/spisok-tablic-iskluchennyh-iz-sbora-statistiki]

c. ***Замечу по ходу****, что в таких ситуациях не стоит надеяться на скорую помощь.* [http://samlib.ru/s/sharafutdinow_r_k/blizko_k_serdcu.shtml]

[9] При поиске использовалось ограничение «-S на расстоянии 1 после», далее выдача была просмотрена и отсортирована вручную. Таким образом, некоторые релевантные примеры, в которых существительное после *по ходу* не является его зависимым, могли не попасть в выдачу. Тем не менее, мы полагаем, что эта неточность не должна повлиять на общее соотношение значений ПХ. Мы полностью согласны с рецензентом «Диалога» о желательности дополнительной корпусной статистики, однако решение этой задачи оставляем на будущее.

(15) *Понтовые Нью-Йоркские пацаны, некисло задвигающие по тамошним клубам, свое дело понимают туго. NY, ~~замечу~~* ***по ходу****, один из самых многонациональных городов в и без того разноцветных Штатах.* [Макар Свирепый. Байда: музыка (2003) // «Хулиган», 2003.12.15]

Интересно, что выражение ‘по ходу дела’ спорадически употребляется также в контекстах, характерных для производных значений выражения ПХ: эпистемического (16) и дискурсивного (17). Мы усматриваем в таких употреблениях следующий путь диахронического развития: *по ходу дела* → *по ходу* (событийное) → *по ходу* (эпистемическое / дискурсивное) → *по ходу дела* (эпистемическое / дискурсивное).

(16) a. *Вашу нетерпимость уже все заметили. И совсем уж от меня лично, не Вам меня учить тому, когда мне заводить ребенка…. У вас психоз* ***по ходу дела****. Никто вас не учит тут!* [Кошки форева! (2008)]

b. *Значит так, уроды с африканы.* ***По ходу дела*** *вы не только жить, но и общаться не умеете. Что ж, флаг вам в руки.* (Интернет-корпус LEEDS [http://corpus.leeds.ac.uk/ruscorpora.html])

c. *Я подпрыгнул на месте. Дэн держал тапочек в руке, брезгливо рассматривая подошву: — Во суки! От соседа,* ***по ходу дела****, лезут…* [Олег Гладов. Любовь стратегического назначения (2000–2003)]

(17) a. [Денис(3d), nick] *Да, форум там есть, но меня там до сих пор не активизировали, а задавать вопросы там без активизации не получается. К тому же я честно написал, что я из Москвы сейчас пишу. Вот если бы я написал, что я из Киева, то меня бы зарегили. Там* ***по ходу дела*** *большинство так и делает.* [Форум: В каких городах сейчас ведется реальное строительство метро? (2007)]

b. *При все при этом диск в общем-то получился скучноватым —* ***по ходу дела****, большинство рок-н-роллов вообще на один мотив написаны, так что результат вполне предсказуемый.* [Х. Ботаник. Байда: Музыка (2004) // «Хулиган», 2004.06.15]

В разделе 3 мы попытаемся придать высказанным здесь интуитивным соображениям о специфике значений ПХ более эксплицитный характер. Однако прежде мы сделаем еще одно важное утверждение о значении ПХ.

2.3. Пресуппозиция

В этом разделе мы покажем, что семантический компонент, привносимый ПХ, имеет пресуппозициональный характер. Чтобы увидеть, что он не составляет часть ассерции, достаточно удостовериться в том, что этот компонент проецируется из-под отрицания.

Проще всего применить этот тест для событийного ПХ (18). Несмотря на известную искусственность такого примера, можно удостовериться, что тот семантический компонент, который, согласно нашему предположению,

составляет семантический вклад ПХ, не подвергается отрицанию. Предложение (18), как и его утвердительный аналог (3b), предполагает наличие фоновых событий, в последовательность которых утверждаемое событие встраивается естественным образом. Под отрицание попадает только наличие этого события в актуальном мире.

(18) *Приятель мой думу думает. Однако неверно, что по ходу он ведёт допрос с пристрастием на предмет анамнеза.*

Важным свойством пресуппозиционального содержания предложения выступает необходимость присутствия в common ground удовлетворяющей его пропозиции или пропозиций (так называемый bridge principle, [Stalnaker 1978])[10]. Именно поэтому пресуппозициональные выражения, возникая в абсолютном начале дискурса, либо вызывают семантическую аномальность, либо подлежат аккомодации. Согласно интуиции носителей, в точности так ведут себя предложения с ПХ. Естественная реакция собеседника на (19a) — запрос на уточнение фоновых событий, как, например, в (19b)[11].

(19) a. — *Привет! Коля по ходу напился.*
b. — *Погоди, а что было-то?*

Похожая реакция наблюдается при употреблении другого типа пресуппозициональных выражений — определенных дескрипций — в начале дискурса. Хотя для русского языка определенный статус многих дескрипций грамматически не маркирован, соответствующий эффект легко выявляется при анализе коммуникативных структур. Еще И. И. Ковтунова [Ковтунова 1976: 42–43] отмечает своеобразную интродуктивную стратегию, при которой «… темой высказывания может быть и неизвестное, новое для читателя или слушателя. Новым нередко бывает начало повествования, главы, абзаца… Приведем примеры. Начало рассказа Чехова «Белолобый»: *Голодная волчиха встала, чтобы идти на охоту.* … Начало романа Ю. Тынянова «Пушкин»: *Майор был скуп. Вздохнув, он заперся у себя в комнате и пересчитал деньги*».

Использование дескрипций ‘голодная волчиха’, ‘майор’ в тематической позиции категорического высказывания первого предложения дискурса вынуждает определенную интерпретацию. Аналогичным образом, определенная

[10] За неимением хорошего переводного эквивалента понятия common ground («the mutually recognized shared information in a situation in which an act of trying to communicate takes place», [Stalnaker 2002: 704]) мы приводим его в исходном виде.

[11] Рецензент «Диалога» пишет: «Трудно спорить об интуиции носителя, но интуиции рецензента такой пример не противоречит (в том идиолекте, где *по ходу* является синонимом *похоже*, такое предложение не удивляет)». Нам трудно судить о том, какие примеры с *похоже* имел в виду рецензент. Если речь идет о замене *по ходу* на *похоже* в (19a), то можно заменить, что такая замена также создает контекст, где необходима аккомодация. Предложение *Вася, похоже, напился* естественно в начале дискурсивной последовательности, если участники коммуникации наблюдают Васю, проявляющего признаки опьянения. Если этого нет, слушающий вынужден аккомодировать по меньшей мере информацию ‘Говорящий наблюдал нечто, что позволило ему заключить, что утверждаемая пропозиция истинна’.

дескрипция на базе личного имени ‘Ваня’ при первом упоминании персонажа в начале рассказа (20), по мысли Е. В. Падучевой, отражает тот факт, что «… Ваня вводится сразу как лицо, известное читателю» [Падучева 1985/2002: 151].

(20) *Купила мать слив и хотела их дать детям после обеда. Они лежали на тарелке. Ваня никогда не ел слив и всё нюхал их.* [Л. Н. Толстой. «Косточка»]

Т. Е. Янко [Янко 2001: 145 и сл.] рассматривает подобную интродуктивную стратегию как «тематизацию вводимого в рассмотрение предмета» и сопоставляет впервые вводимым референтам с тематическим акцентом (21a) сложную коммуникативную структуру с анафорически связанными элементами (21b), подвергающуюся компрессии (обсуждение этого анализа см. также в [Падучева 2012, 2016]).

(21) a. *Утка↗ плавала по реке↘.*
b. *Была утка↘. Она* (эта утка↗) *плавала по реке↘.*

На наш взгляд, перифраз (21b) отражает процесс аккомодации пресуппозиции: тематическая позиция в интродуктивном предложении вынуждает определенную интерпретацию именной группы ‘утка’, но в отсутствие в common ground активированного референта высказывание (21a) вынуждает слушающего добавить соответствующую экзистенциальную пропозицию (‘была утка’).

В отличие от определенной дескрипции, однако, пресуппозиция, вводимая ПХ, сопротивляется аккомодации в несколько большей степени. Мы связываем это с тем, что в пресуппозицию входит не только экзистенциальное утверждение о наличии фоновых событий, но и их контекстно доступные дескриптивные свойства. Подробнее об этом будет сказано в разделе 3.

Сформулировав эти предварительные замечания, в следующем разделе мы попытаемся дать трем значениям ПХ более эксплицитную характеристику.

3. Наброски анализа

Ниже представлены теоретико-модельные построения, позволяющие, как мы надеемся, представить обобщения из прошлого раздела в более явном виде и внести, где это потребуется, необходимые уточнения. В этом же разделе собраны воедино соображения о семантической эволюции ПХ.

3.1. Событийное ПХ

Мы предлагаем следующий теоретико-модельный анализ для событийного ПХ:

(22) $|| ПХ_{EV} || = \lambda P.\lambda e\text{:}\ \exists e_1, \ldots, e_n\ [Q_1(e_1) \wedge \ldots \wedge Q_n(e_n) \wedge \tau(\bigoplus_{i=1}^{n} e_i) \otimes \tau(e) \wedge NU_C (\bigoplus_{i=1}^{n} e_i \oplus e)].\ P(e)$

Согласно (22), событийное ПХ представляет собой функцию эквивалентности с пресуппозицией. Оно применяется к событийному предикату и возвращает тот же событийный предикат при условии, что событие из его экстенсионала удовлетворяет этой пресуппозиции. В противном случае значение выражения с ПХ не определено. Этот анализ отражает обсуждавшуюся в предыдущем разделе интуицию, согласно которой семантическое содержание ПХ имеет пресуппозициональный характер.

Рассмотрим подробнее элементы, из которых складывается (22). Пресуппозиция (22) — это утверждение, что в текущем мире имеет место одно или более событие e_1, …, e_n, которое мы выше назвали фоновым. Каждое событие удовлетворяет собственной событийной дескрипции Q_1, …, Q_n. Дескрипции представлены в (22) как несвязанные переменные, получающие значение в результате оценки переменных. Это делает дескриптивные свойства фоновых событий контекстно-зависимыми.

В диахронически первичной конструкции вида «ПХ N_{GEN}» (*по ходу дела, по ходу фильма, по ходу заседания* и т.п., см. примеры в разделе 2.2) дескрипция фонового события представляет собой аргумент ПХ, а не элемент контекста.

Процесс диахронического развития, соответственно, следует описывать как утрату синтаксической валентности на именную группу с событийной референцией[12]. Семантическая валентность сообразно этому претерпевает изменения: реализующие ее элементы приобретают пресуппозициональный характер, ограничение на единственность дескрипции для фоновой ситуации снимается, сами фоновые ситуации приобретают экзистенциальную интерпретацию.

Мы предполагаем наличие в пресуппозиции контекстно-зависимых событийных дескрипций Q и, соответственно, отклоняем альтернативный анализ в (23) ввиду простого соображения. (23) отличается от (22) тем, что пресуппозиция утверждает существование фоновых событий, но полностью умалчивает об их дескриптивных свойствах.

$$(23)\ ||\ ПХ_{EV}\ || = \lambda P.\lambda e\text{: } \exists e_1,\ldots,e_n\ [\tau(\bigoplus_{i=1}^{n} e_i) \otimes \tau(e) \wedge NU_C\ (\bigoplus_{i=1}^{n} e_i \oplus e)].\ P(e)$$

Основное соображение, которое заставляет нас предпочесть (22), а не (23), связано с описанным в предыдущем разделе наблюдением: предложения с ПХ в пустом контексте более аномальны, чем определенные дескрипции. Пресуппозиция существования, которая обычно предполагается у определенных дескрипций (хотя см. аргументы против этого анализа в [Coppock, Beaver 2015]) и пресуппозиция существования в (23) должны проявлять одинаковые

[12] В контексте текущего обсуждения мы предпочитаем говорить о диахроническом развитии, а не о грамматикализации в более узком смысле. У *по ходу* более или менее отчетливо просматривается характерные диахронические изменения в области семантики — расширение значения за счет конвенционализации импликатур и метонимический перенос в другой домен. Мы оставляем открытым для будущего исследования вопрос о сопутствующих изменениях плана выражения (фонологической редукции, изменении просодического статуса и т. п.), которые в совокупности с семантической эволюцией позволили бы говорить о грамматикализации в узком смысле. Мы признательны анонимному рецензенту, замечание которого побудило нас к этому уточнению.

возможности в плане аккомодации. Если приемлемость пресуппозициональных выражений в пустом контексте ограничена именно возможностью аккомодации, можно было бы ожидать, что ПХ по меньшей мере так же приемлемо, как определенные дескрипции. Однако это не так.

Предположив в (22) несвязанные переменные, пробегающие по событийным дескрипциям, мы получаем возможность объяснить эту асимметрию. Несвязанные переменные, такие, например, как анафорические местоимения, получают интерпретацию посредством функций оценки переменных. Чем беднее контекст, тем более труднодоступны функции, определенные для индекса данной переменной. Мы тем самым связываем невозможность аккомодации пресуппозиции в (22) с аналогичной невозможностью понять в пустом контексте, на кого указывает местоимение в предложении типа Он пришел. Говоря неформально, чтобы воспринять предложение с ПХ как приемлемое, слушающий должен не только принять к сведению существование фоновых событий, но и получить информацию о том, каковы в точности эти события. Если информации нет, что и происходит, как правило в пустом контексте, ПХ вызывает аномальность.

Следующий, центральный компонент семантики ПХ требует пересечения времени утверждаемого события e, τ(e), и времени осуществления фоновых событий $\tau(\bigoplus_{i=1}^{n} e_i)$.

Этот компонент описывает самую отчетливо воспринимаемую характеристику ПХ — одновременность тех и других. В (22) мы используем не отношение идентичности (“=”), гарантирующее одновременность, а более слабое отношение временного пересечения (“$\otimes_T$”) в виду примеров типа (24). В (24) очевидно не требуется, чтобы строительство столицы началось одновременно с началом самой ранней войны, а закончилось в момент завершении самой поздней. Более того, в строительстве возможны остановки, не сопровождающиеся одновременными перерывами в военных действиях. Отношение пересечения, в отличие от отношения идентичности, допускает весь этот диапазон возможностей.

(24) *Хан вел четыре войны с соседями, а по ходу продолжал отстраивать свою столицу.*

Наконец, последний компонент, который мы выше обозначили как событийную когерентность, мы даем в (22) в виде неанализируемой функции NU_C («контекстно-зависимая естественная единица»). Не прорабатывая технические детали, опишем действие этой функции. Она применяется к событию, удостоверяется, что событие состоит из нескольких атомарных частей и возвращает значение “1”, если эти части образуют естественную совокупность. Естественность совокупности вычисляется применительно к текущему контексту C. Мы оставляем без дальнейшей экспликации это ключевое понятие. Как и многие лингвистические понятия, в названии которых фигурирует слово «естественный» (естественный класс, естественный ход событий и т. п.), оно может быть осмыслено более чем одним способом. Естественность некоторой совокупности S, относящейся к домену объектов D, $S \subseteq D$, можно понимать как наличие у ее элементов (возможно, очень контекстно-зависимого) свойства,

которого лишены все прочие элементы, D\S, и как отсутствие у элементов D несовместимых свойств в противоположность D\S. Можно также рассматривать естественный класс событий как единое макрособытие, а элементы класса — как его подсобытия. В этом случае, по-видимому, становится возможным теоретико-вероятностное описание: осуществление одного из элементов S влияет на вероятность осуществления других элементов S в большей степени, чем на вероятность осуществления элементов D\S. Оставляя этот вопрос открытым, укажем лишь, что присутствие этого компонента в семантике ПХ позволяет не только описать интуицию по поводу примеров (4)–(5), о которых шла речь выше, но и предложить правдоподобный сценарий перехода от событийного ПХ к эпистемическому и дискурсивному. Об этом будет сказано ниже.

3.2. Эпистемическое ПХ

Для эпистемического ПХ мы предлагаем семантику в (25).

(25) $|| ПХ_{\text{EPIST}} || = \lambda P.\lambda e$: $\exists e_1, \ldots, e_n\, [Q_1(e_1) \wedge \ldots \wedge Q_n(e_n) \wedge \text{infer}_c(\exists e[e=(\bigoplus_{i=1}^{n} e_i)], \exists e\, P(e))].\ P(e)$

В (25) вместо компонентов, отвечающих за временное пересечение $(\tau(\bigoplus_{i=1}^{n} e_i) \otimes \tau(e))$ и естественность класса $NU_C\ (\bigoplus_{i=1}^{n} e_i \oplus e)$, представлен компонент $\text{INFER}_C\ (\exists e[e=(\bigoplus_{i=1}^{n} e_i)], \exists e\ P(e))$, который отвечает за **контекстную инференцию**. В соответствии с (25), пропозиция ∃e P(e), описывающая наличие в мире утверждаемого события, представляет собой результат логического вывода, опирающегося на пропозицию, описывающую наличие фоновых событий. Говоря неформально, (7b), например, осмысленно ровно в том случае, когда утверждение о том, что ‘Ты в других регионах не был’, подкрепляется наличием в актуальном мире фонового события ‘Собеседник высказал суждение «Подмосковье — самый богатый регион в заМКАДье»’.

Рассмотрим подробнее частные случаи эпистемического употребления. Выясняется, что характер логического вывода зависит от взаиморасположения утверждаемого и фоновых событий во времени. (Напомним, их временное пересечение при эпистемическом употреблении допускается, но не требуется.)

Первый случай: утверждаемое событие предшествует фоновым. Эта возможность иллюстрируется в (7d), повторяемом как (26):

(26) *Слушай, а в правом <ботинке> еще вода. Ты, по ходу, не все вылил.*
[Андрей Геласимов. Год обмана (2003)]

В этом случае логический вывод имеет характер абдукции. **Абдукция** (термин предложен Ч. Пирсом), или **вывод наилучшего объяснения**, — это разновидность вывода, которую в первом приближении можно описать следующим образом:

(27) *При наличии наблюдаемого положения вещей E и возможных объяснений $H_1, \ldots H_n$, следует заключить, что истинно то H_i, которое объясняет E наилучшим образом.*

Абдукция — важнейшая часть обыденной логики. Эту операцию, например, проделывает врач, когда, изучив симптомы заболевания, ставит диагноз. Абдукция широчайше представлена в естественнонаучных рассуждениях, когда требуется дать каузальное объяснение наблюдаемому явлению, а в распоряжении исследователя имеется более одной гипотезы.

В (26) пропозиция ‘ты еще не все вылил’, соответствующая утверждаемому событию, абдуцируется на основании пропозиции ‘в левом ботинке вода’, описывающей фоновое событие. Эта пропозиция и подается как оптимальное объяснение наблюдаемого положения дел.

Условие успешной абдукции для эпистемических употреблений ПХ можно сформулировать следующим образом:

(28) *Пусть Common Ground — это множество пропозиций, которые участники коммуникации рассматривают как истинные в текущем контексте C. Пусть q — пропозиция, описывающая фоновые события* (в (25) — $\exists e[e=(\bigoplus_{i=1}^{n} e_i)]$). *Пусть p, $p \notin CG$, — пропозиция, описывающая утверждаемое событие* (в (25) — $\exists e\, P(e)$), *причем p дает каузальное объяснение q. Тогда*

a. *q не вытекает из common ground без p:* $CG \not|^{1} q$;

b. *p в совокупности с common ground достаточна для вывода q:* $CG \dot{\cup} \{p\} \models q$;

c. *p — наиболее вероятное каузальное объяснение:*

$\forall p'\ [CG \dot{\cup} \{p'\} \models q \rightarrow [\ [p' \text{ cause } q] <_{\text{likely}} [p \text{ cause } q]]$.

(28a) и (28b) самоочевидны. (28a) обеспечивает, чтобы утверждаемая пропозиция была необходима для объяснения наблюдаемого положения дел *q*. (28b) вводит условие, чтобы эта пропозиция в совокупности с общей информационной базой была достаточна для объяснения *q*. (28c) требует, чтобы пропозиция была наиболее вероятной причиной *q*.

Такая семантика делает соответствующие употребления эпистемического ПХ похожими на многие другие выражения, использующиеся в абдуктивных контекстах — *я вижу, оказывается*, инферентивные перфекты в тех языках, где перфект грамматикализован как категория, выражающая значения непрямой засвидетельствованности. Абдуктивное ПХ в примерах типа (26), как кажется, не имеет принципиальных отличий от показателей такого рода и может анализироваться аналогично, то есть как в (28) (см. [Татевосов 2018]). Сходство эпистемического ПХ и целого ряда других выражений, которое иллюстрируется в (8), получает тем самым последовательное объяснение.

Характер логического вывода меняется, если фоновые события предшествуют утверждаемому. Событие *e*, следующее во времени за *e′*, нельзя использовать для причинного объяснения *e′*. Соответственно, в такой временной конфигурации вместо абдукции возникает более привычная дедукция. Из истинности посылок, которые представляют собой описание фоновых ситуаций ($\exists e[e=\bigoplus_{i=1}^{n} e_i]$) в комбинации с информацией, содержащейся в common ground, выводится утверждаемая пропозиция $\exists y\, P(e)$:

(29) $CG \cup \{q\} \models p$

Пример такой возможности — (1b), повторяемый как (30):

(30) [Roberta, жен] *И надо ехать на дачу, закруглять» стройку «, а там по агентурным данным» конь не валялся «… и придёт трындец моему так нелегко добытому спокойствию :* (Короче, деньги кончились, кухня походу только в следующем году :)*]* # [Форум: Апгрейд дачной кухни (2011–2013)]

Элемент *q* в этом случае — утверждение о наличии в актуальном мире суммы ситуаций ‘на стройке конь не валялся’ и ‘деньги кончились’, а элементы CG — общие знания типа ‘строительство объекта без денег невозможно’. Из всего этого с неизбежностью вытекает *p* — утверждение ‘кухня (будет построена) только в следующем году’.

Похожее переключение абдукции и дедукции, связанное с изменением временного плана описываемой ситуации, опять же не уникально. В [Татевосов 2017] описывается содержательно близкий процесс, происходящий с формой адмиратива в татарском языке.

Наконец, третья логическая возможность при эпистемическом ПХ соответствует временно́му пересечению утверждаемого и фонового событий (то есть темпоральному отношению, единственно возможному для событийного ПХ). Как представляется, в этом случае эпистемическое ПХ допускает и абдуктивный и дедуктивный логический вывод. Эти две возможности иллюстрируются в (31)–(32):

(31) *Салон не потерся, пластик в салоне не гремит и не скрипит* (мечта владельцев отечественных ТАЗиков:)*), шумоизоляция неплохая, кондиционер работает. По ходу, где-то сифонит глушак, шума стало многовато.* [Внедорожник для хулигана: Terrano II (2004) // «Хулиган», 2004.07.15]

(32) — *Он наверно в бильярдах потерялся.*
— *Он писал, что с бильярдом завязывает и на покер переходит — так что походу не в бильярдах он, а в покерах.*
[https://tradernet.com/feed/postId/1088860]

В (31) утверждение ‘сифонит глушак’ возникает в результате абдукции и выступает объяснением фоновой ситуации ‘шума стало многовато’. В (32), напротив, утверждение ‘не в бильярдах он, а в покерах’ подается как следствие, вытекающее из наблюдения, изложенного в фоновом предложении (*Он писал, что с бильярдом завязывает и на покер переходит*). Абдуктивное прочтение, при котором утверждаемое предложение вводится как объяснение фонового, в этом контексте, по-видимому, невозможно.

Несколько соображений о семантическом переходе, связывающем эпистемическое и событийное ПХ. В рафинированном виде содержание этого перехода можно представить так, как показано в (33):

(33) $\alpha \otimes_{T} \beta \wedge NU_{C}(\alpha,\beta) \Rightarrow INFER_{C}(\alpha, \beta)$

Согласно (33), две сущности α и β, пересекающиеся во времени и образующие некоторое целостное единство, переосмысливаются как связанные

причинно-следственным отношением. Это позволяет использовать одну из них, чтобы составить суждение о другой.

Мы предполагаем, что именно в (33) дает о себе знать компонент, который мы неформально называем событийная когерентность и за который отвечает функция NU_C. Ключевой элемент нашего рассуждения — то соображение, что отношение временно́го пересечения слишком общо, чтобы самостоятельно обеспечить необходимый переход. Диахронической семантике известны случаи, когда временные отношения реинтерпретируются как каузальные (см., например, [Wegener 2002]). Однако существенное условие такой реинтерпретации состоит в том, что отношение однозначно задает взаиморасположение событий во времени. Идентифицировать одно событие как причину другого возможно только в том случае, когда известно, что они либо следуют друг за другом во времени, либо развертываются одновременно. Отношение пересечения недостаточно специфицировано, чтобы удовлетворить этому условию.

Наша гипотеза состоит в том, что событийная когерентность — тот семантический элемент, который восполняет этот недостаток отношения пересечения. Благодаря событийной когерентности в содержание сообщения вводится информация, что утверждаемое и фоновое события образуют естественное целое (то есть, например, что осуществление одного события влияет на вероятность осуществления другого). С помощью известных максим речевой коммуникации из этого выводятся каузальные импликатуры, которые делают возможным переход в (33).

3.3. Дискурсивное ПХ

Как мы уже сказали, наше предположение по поводу дискурсивного ПХ состоит в том, что его значение полностью идентично событийному ПХ, однако реализуется в другом домене — домене речевых актов. Все элементы пресуппозиции в (22) соответствующим образом реинтерпретируются.

Первый элемент, $Q_1(e_1) \wedge \ldots \wedge Q_n(e_n)$, в результате реинтерпретации становится конъюнкцией совершившихся в актуальном контексте речевых актов. Элемент, ответственный за временное пересечение, $\tau(\bigoplus_{i=1}^{n} e_i) \otimes \tau(e)$, теперь указывает на временное пересечение речевых актов. Событийная когерентность, $NU_C(\bigoplus_{i=1}^{n} e_i \oplus e)$, становится дискурсивной когерентностью.

Среди семантических теорий речевых актов нам известна одна, которая позволяет реализовать эту идею самым непосредственным образом. М. Крифка [Krifka 2014] предлагает динамический анализ речевых актов, опирающийся на идею, что совершение речевого акта представляет собой событие, определенным образом изменяющее мир. «Речевые акты, — пишет М. Крифка, — не могут быть истинны или ложны в том или ином мире в том или иной момент. Они создают новые факты, меняющие мир. Коммуникация не просто меняет common ground участников коммуникации — она меняет мир как таковой».

При таком понимании речевых актов становится возможным рассматривать их как события — и в этом качестве подавать в семантическом описании. В [Krifka 2014] этот шаг сделан опосредованно: семантические операторы, описывающие речевые акты, определяются как функции, в качестве аргументов принимающие индексы, а не события. События моделируются опосредованно — в терминах разных истинностных значений в разных индексах (индекс — это точечный объект, который представляет собой комбинацию возможного мира и момента времени). Примерно так Д. Даути задает темпоральную семантику для своих аспектуальных классов [Dowty 1979]: скажем, истинность пропозиций, построенных с помощью предиката класса достижений (например, ‘сломаться’) определяются через истинность пропозиции ‘быть целым’ на интервале *t* и ложность этой пропозиции на следующем за ним интервале *t′*.. Точно так же, событие речевого акта можно понимать как изменение истинностного значения пропозиции ‘У говорящего есть социальные обязательства по поводу истинности пропозиции’ при переходе от одного индекса к другому. Поэтому, как отмечает сам М. Крифка, введение эксплицитных событийных переменных в семантические представления, предполагаемые его анализом, — задача сугубо техническая. В пределах этих заметок мы не будем развертывать достаточно громоздкую машинерию, стоящую за теорией М. Крифки, и отсылаем читателя за подробностями к первоисточнику. Для текущего обсуждения существенно одно: если верно, что речевые акты — это особый род событий, мы получаем естественное основание для распространения событийной семантики ПХ в дискурсивную сферу. Параллелизм событийных и дискурсивных употреблений ПХ получает простое и наглядное объяснение.

Подводя итог этого раздела, суммируем сказанное выше о соотношении трех типов употребления ПХ. Предложенные нами обобщения позволяют выстроить достаточно простую картину. Исходным является событийное ПХ, семантика которого показана в (22). Из событийного ПХ независимым образом возникают эпистемическое и дискурсивное. Первое представляет собой результат реинтерпретации двух компонентов: временного пересечения и событийной когерентности. В результате реинтерпретации между утверждаемым и фоновым событиями устанавливается связь, опирающаяся на логический вывод — либо абдуктивный, либо дедуктивный. Второе создается переносом семантики событийного ПХ в другую область — область речевых актов. Естественным основанием для такого переноса, представляющего собой, по-видимому, род метонимии, выступает возможность конституирования речевых актов как особого рода событий.

4. Заключение

Завершая эту статью, мы готовы предложить ответы на вопросы, сформулированные в самом начале. Вклад ПХ в интерпретацию имеет пресуппозициональный характер. Пресуппозиция ПХ предъявляет определенные требования к языковому контексту и common ground и, если эти требования

удовлетворяются, выдает пропозиции, выступающей ее сферой действия, лицензию на дальнейшее существование в дискурсе[13].

В зависимости от содержания пресуппозиции выделяется три частных значения ПХ: событийное, дискурсивное и эпистемическое. В событийном значении пресуппозиция ПХ складывается из трех компонентов: в текущем мире должны иметь место фоновые события, они должны иметь временное пересечение с утверждаемым событием, утверждаемое и фоновые события должны быть когерентны. Дискурсивное ПХ получается переносом этой же семантики в область речевых актов. Эпистемическое ПХ связано с событийным посредством инферентивной реинтерпретации, когда временное пересечение и событийная когерентность позволяют осмыслить фоновые ситуации как эпистемическое основание для высказывания об утверждаемой.

Предполагаемый путь диахронического развития ПХ показан в (34):

(34) по [ходу NP_{GEN}] ⟶ ПХ *pro* ⟶ $ПХ_{EV}$ ⟶ $ПХ_{EPIST}$, $ПХ_{DIS}$

Первым шагом стала утрата словом *ход* валентности на генитивную именную группу. Эта утрата, по-видимому, проходила через стадию реализации актанта непроизносимым местоимением *pro*. Следующей стадией стало появление событийного ПХ, которое далее независимо развивалось в направлении дискурсивного и эпистемического ПХ.

Литература

1. *Beaver D.* (2001), Presupposition and Assertion in Dynamic Semantics, Stanford: CSLI Publications.
2. *Coppock E., Beaver D.* (2015), Definiteness and determinacy, Linguistics and Philosophy, vol. 38, pp. 377–435.
3. *Dowty D. R.* (1979), Word Meaning and Montague Grammar, Dordrecht: Reidel.
4. *Heim I.* (1983), On the projection problem for presuppositions, Second Annual West Coast Conference on Formal Linguistics, Stanford, pp. 114–126.
5. *Kovtunova I. I.* (1976), Modern standard Russian: word order and information structure [Sovremennyj russkij jazyk: Porjadok slov i aktual'noe chlenenie predlozhenija], Moscow: Nauka.

[13] Мы оставляем в стороне вопрос о синтаксической дистрибуции различных употреблений ПХ, которая требует дополнительного исследования. Из априорных соображений можно предположить, что событийное ПХ присоединяется в той же позиции, что и обстоятельства образа действия; а эпистемичекое и дискурсивное — там же, где эвиденциальные обстоятельства и обстоятельства, ориентированные на ситуацию речевого акта. В терминах известной иерархии Г. Чинкве это соответственно VoiceP, $Mood_{Evidential}$P и $Mood_{SpeechAct}$P. Сформулировать более точные обобщения мешает то, что в русском языке поверхностное линейное расположение обстоятельства мало помогает идентификации его синтаксического класса; существенно более информативно попарное сравнения обстоятельств разных типов. В случае ПХ это задача на будущее. Мы признательны анонимному рецензенту, замечание которого побудило нас к этому уточнению.

6. *Krifka M.* (2014), Embedding illocutionary acts, Recursion: Complexity in cognition, Berlin: Springer, pp. 59–87.
7. *Krongauz M. A.* (2008), Russian on the verge of a nervous breakdown [Russkij jazyk na grani nervnogo sryva], Moscow.
8. *Ozhegov S. I., Shvedova N. Y.* (2006), Dictionary of Russian [Tolkovyj slovar' russkogo jazyka], 4th edition, Moscow: A TEMP.
9. *Nikitina T. G.* (2003), Youth slang. Dictionary [Molodezhnyj sleng. Tolkovyj slovar'], Moscow: Astrel'.
10. *Paducheva E. V.* (1985/2002), Utterance and its relatedness to the reality [Vyskazyvanie i ego sootnoshenije s dejstvitel'nost'ju], Moscow: URSS
11. *Paducheva E. V.* (2012), On the semantics of information structure: non-derived structure, and linear-accentual transformations [K semantike kommunikativnoj struktury: ishodnye struktury i linejno-akcentnye preobrazovanija], Computational Linguistics and Intellectual Technologies. Vol. 11. Proceedings of the International Conference "Dialog 2012" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii. Vyp. 11. Po materialam ezhegodnoj mezhdunarodnoy konferentsii "Dialog 2012"], Moscow, pp. 522–533.
12. *Paducheva E. V.* (2016), Information structure and linear-accentual transformations of a sentence (evidence from Russian) [Kommunikativnaja struktura i linejno-akcentnye preobrazovanija predlozhenija (na materiale russkogo jazyka)], Clause architecture in parametric models [Arhitektura klauzy v parametricheskih modeljah], Moscow, YSK, pp. 25–75.
13. *Stalnaker R.* (1978), Assertion, Syntax and Semantics, vol. 9, pp. 315–332.
14. *Stalnaker R.* (2002), Common Ground, Linguistics and Philosophy, vol. 25, pp. 701–721.
15. *Tatevosov S. G.* (2017), Perfect, evidentiality and mirativity [Perfekt, evidentsialnost' i admirative], Elements of Tatar in a typological perspective: Mishar dialect [Elementy tatarskogo jazyka v tipologicheskom osveshchenii: Misharskij dialekt], Moscow: BukiVedi, pp. 124–145.
16. *Tatevosov S. G.* (2018), Evidentiality and abduction [Evidentsial'nost' I abduktsija], ms., Lomonosov Moscow State University.
17. *Ushakov D. N.* (2001), Dictionary of Russian [Tolkovyj slovar' russkogo jazyka], vol. 3, Moscow: Veche.
18. *Wegener H.* (2002), The evolution of the German modal particle denn, New Reflections on Grammaticalization, Amsterdam: Benjamins, pp. 379–394.
19. *Yanko T. E.* (2001), Communicative strategies of Russian [Kommunikativnye strategii russkoj rechi]. M.: YSK, 2001.

LEVERAGING DEEP NEURAL NETWORKS AND SEMANTIC SIMILARITY MEASURES FOR MEDICAL CONCEPT NORMALISATION IN USER REVIEWS

Miftahutdinov Z. (zulfatmi@gmail.com),
Tutubalina E. (elvtutubalina@kpfu.ru)

Kazan Federal University, Kazan, Russia

Nowadays a new yet powerful tool for drug repurposing and hypothesis generation emerged. Text mining of different domains like scientific libraries or social media has proven to be reliable in that application. One particular task in that area is medical concept normalization, i.e. mapping a disease mention to a concept in a controlled vocabulary, like Unified Medical Language System (UMLS). This task is challenging due to the differences in language of health care professionals and social media users. To bridge this gap, we developed end-to-end architectures based on bidirectional Long Short-Term Memory and Gated Recurrent Units. In addition, we combined an attention mechanism with our model. We have done an exploratory study on hyperparameters of proposed architectures and compared them with the effective baseline for classification based on convolutional neural networks. A qualitative examination of the mentions in user reviews dataset collected from popular online health information platforms as well as quantitative one both show improvements in the semantic representation of health-related expressions in user reviews about drugs.

Key words: medical concept mapping, medical concept normalization, deep learning, UMLS, recurrent neural networks, information extraction

1. Introduction

There were many novel applications of Natural Language Processing (NLP) to biomedical information in recent years. Most of researchers' attention attracts task of Named Entity Recognition (NER). Many applications of NER have been applied to scientific literature and electronic health records. And comparatively little work was carried out on social media texts of individuals undergoing medical treatment.

Social media in recent years had become a virtually inexhaustible source of people's opinions on the wide variety of topics. In this work, our focus is patients' opinions on drug effects, i.e. patients' reports. Progressive improvement of text mining approaches applied to patient reports in social media by the terms of accuracy and recall has multiplicative effect on several areas including pharmacovigilance (especially, for new drugs), drug repurposing, and understanding drug effects in the context of important and yet not well studied other factors such as concurrent use of other drugs, diet, and lifestyle.

We study the patients' comments on social media in an aspect of discovering disease-related medical concepts from. In the context of this problem, we map a text written in the informal language of social media (e.g. "I can't fall asleep all night" or "head spinning a little") to formal medical language (e.g. "insomnia" and "dizziness" respectively).

This goes beyond simple straightforward matching of natural language expressions with vocabulary elements: string matching approaches may not be able to link the social media language to the medical concepts due to few or an absence of overlapping words. We call the task of mapping everyday life language to medical terminology medical concept normalization (or medical concept mapping). The main benefit of solving this task is bridging the gap between the language of lay public and medical professionals.

The described task seems to be uneasy since patients post in social media texts on different illness concepts (a wide variety of one's from conditions like major depressive disorder to informal phrases describing specific symptoms such as "woke up too early" or "mucus building up in my lungs") and a wide diversity of drug reactions (e.g., "excessive sweating at night", "slept like a baby", or "clearing up an infection"). Also, we should mention that the data from social networks typically contain a lot of noise such as typos, misspellings, incorrect grammar, hashtags, abbreviations, and different variations of the same word.

Formally speaking, this task is related to several well-known NLP challenges including paraphrase detection, word sense disambiguation, and entity linking where an entity mention is mapped to a unique concept in an ontology after solving the disambiguation problem [1, 2]. In recent studies, there were proposed some approaches to this challenge treating the task of linking a one- or multi-word expression to a knowledge base as a supervised sequence labeling problem. Miftahutdinov and Tutubalina [3] proposed an encoder-decoder model based on bidirectional recurrent neural networks (RNNs) to translate a sequence of words from a death certificate into a sequence of medical codes. Two recent works present similar approaches [4, 5] that utilize RNNs for normalization of tweets' phrases at the AMIA 2017 Social Media Mining for Health Applications workshop, while Limsopatham and Collier [6] experimented with convolutional neural networks (CNNs) on social media data. These works demonstrated usage of deep learning techniques for medical concept normalization. In this paper, we experimented with more complex RNN architectures with

an attention mechanism and additional linguistic knowledge. Moreover, we study the impact of different word embeddings. We conduct extensive experiments on a real-life dataset from Askapatient.com and demonstrate the effectiveness of the proposed method for medical concept mapping.

2. Background

The most popular knowledge-based system for mapping texts from scientific literature and clinical records to medical identifiers are MetaMap [7] and DNorm [8]. MetaMap was developed by the National Library of Medicine (NLM) in 2001 and has become a de-facto baseline method for many recent studies. This system is based on UMLS and a linguistic approach using lexical lookup and variants by associating a score with phrases in a sentence. Leaman et al. introduced a DNorm system for assigning disease mentions from PubMed abstracts a unique identifier from a MEDIC vocabulary, which combines terminology from Medical Subject Headings (MeSH) and Online Mendelian Inheritance in Man (OMIM) [8]. DNorm consists of a text processing pipeline, including the named entity recognizer to locate diseases in the text, and a normalisation method. The normalisation method is based on a pairwise learning-to-rank technique using the tokens from all mentions as features. DNorm outperformed MetaMap as the baseline.

While there has been a lot of work on named entity recognition from social media posts that has been done over the past 7 years [5, 9, 10, 11, 12, 13, 14, 15, 16], relatively few researchers have looked at assigning social media phrases to medical identifiers. First Social Media Mining shared task workshop (organized as part of the Pacific Symp. on Biocomputing 2016) was designed to mining pharmacological and medical information from social media, with a competition based on a published dataset [13]. Task 3 is devoted to medical concept normalisation, where participants were required to identify the UMLS concept for a given ADR. The evaluation set consisted of 476 ADR instances. Sarker et al. [13] noted that there had been no prior work on normalisation of concepts expressed in social media texts, and task 3 did not attract much attention from the researchers.

Recently, two teams namely UKNLP [4] and gnTeam [5] participated in the Second Social Media Mining for Health (SMM4H) Shared Task and submitted their systems for automatic normalisation of ADR mentions to MedDRA concepts. For the task 3, Sarker et al. [17] created a new dataset of tweets' phrases. The training set for this task contains 6,650 phrases mapped to 472 concepts, while the testing set consisted of 2,500 phrases mapped to 254 classes. We also note that organizers of this task did not describe the corpus creation in details as well as not providing corpus statistics, e.g., the overlap percentage between training and testing sets. Teams' systems showed similar results. The gnTeam's approach contained three components for preprocessing and classification. The first two components corrected spelling mistakes and converted sentences into vector-space representation, respectively. For the third step, GnTeam adopted multinomial logistic regression model which achieved the accuracy of 0.877, while the bidirectional GRU achieved the accuracy of 0.855. As input, the network adopted the GoogleNews embeddings trained on a Google News corpus

due to higher results the highest performance over embeddings trained on tweets. The ensemble of both classifiers showed slightly better performance and achieved the accuracy of 0.885. The UKNLP's system adopted hierarchical LSTM in which a phrase is segmented into words and each word is segmented into characters. Word embeddings were trained on a Twitter corpus. Hierarchical Char-LSTM achieved the accuracy of 0.872, while hierarchical Char-CNN performed slightly better and achieved the accuracy of 0.877. We note this corpus of tweets for future work since the official test data is available for the shared task participants only by the time of publication.

Recently, Limsopatham and Collier [6] experimented with Convolutional Neural Networks (CNN) and pre-trained word embeddings for mapping social media texts to medical concepts. For evaluation, three different datasets were used. The authors created two datasets with 201 and 1,436 Twitter phrases which mapped to concepts from a SIDER database. The third dataset is the CSIRO Adverse Drug Event Corpus (CADEC) [2] which consists of user reviews from askapatient.com. The authors observed that training can be effectively achieved at 40–70 epochs. As input, the network concatenated embeddings of words. The GoogleNews embeddings improved results significantly over embeddings on medical articles. Experiments showed that CNN (accuracy 81%) outperformed DNorm (accuracy 73%), RNN (accuracy 80%) and a multi-class logistic regression (accuracy 77%) on the AskAPatient corpus (as well as corpora of tweets). This work is the closest to ours in the use of deep learning technology and semantic representation of words. However, we found that only approximately 40% of expressions in the test data are unique, while the rest of expressions occur in the training data. Therefore, the presented accuracy may be too optimistic. We believe that future research should focus on developing extrinsic test sets for medical concept normalisation.

3. Methods

In this section, we will discuss major challenges in this task and applied neural architectures.

3.1. Recognition of Different Word Variances

The task of medical concept normalisation is closely related to the problem of word sense disambiguation and terminological variance. There are major challenges which disease mention recognition methods as well as term extraction methods face:

(i) exical, morphological, and syntactic variants;
(ii) paraphrases, synonyms;
(iii) abbreviations;
(iv) ambiguity;
(v) misspellings.

The examples of three-form phrases from the CADEC corpus are presented in Table 1.

Table 1. Examples of three-form phrases with corresponding medical concepts

Free-form Phrases	Medical concept	SNOMED ID
lower pelvic pain	Pain in pelvis	30473006
uterus contractions	Uterine spasm	29542008
something wrong with my uterus	Uterus problem	289621007
stomach issues	Stomach problem	300306001
slightly heavier menstrual cycle	Menorrhagia	386692008
inflammation in my back muscles	Muscle cramp	55300003
inflammation in my neck	Cervical arthritis	387801000
heavy menstrual bleeding	Menorrhagia	386692008
acidic bile in my mouth	Acid reflux	698065002
could only walk less than 100 meters	Reduced mobility	8510008
very painful joints	Arthralgia	57676002
starting to upset my stomach	Stomach ache	271681002
can't sleep	Insomnia	193462001
high BP	Increased venous pressure	69791001
pulse is still extremely high	Pulse fast	86651002

3.2. Proposed Model for Concept Mapping

We propose a deep approach for mapping entity mentions to medical codes. We first convert each mention into a semantic representative vector using bidirectional LSTM or GRU [18–23] with attention mechanism on top of the embedding layer. We use the hyperbolic tangent as activation function. Then, a set of features are extracted using the cosine similarity between mentions and medical concepts from the UMLS Metathesaurus. For model training, we use the cross-entropy error between gold distribution and predicted distribution as the loss function. The model is depicted in Fig. 1.

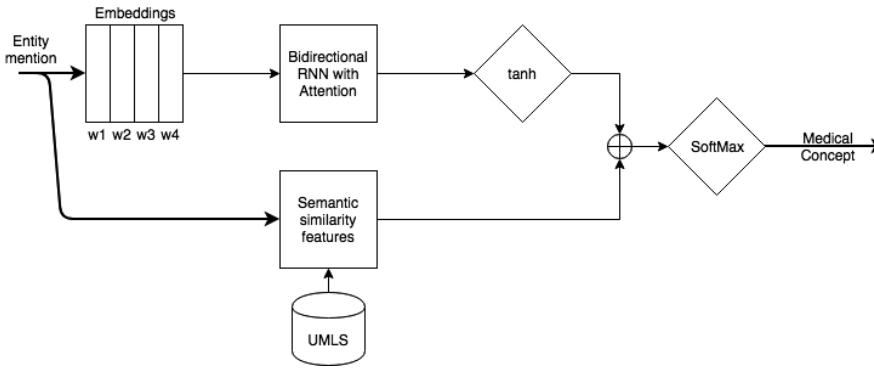


Figure 1: Proposed architecture for medical concept normalization

3.3. Semantic Similarity Features

We extract a set of features to enhance the representation of the phrases. These features consist of cosine similarity between the vectors of the input phrase and a concept in a medical terminology dictionary. This dictionary includes medical codes and synonyms from the UMLS Metathesaurus (version 2017 AA), where codes are presented in the CADEC corpus. We apply the following strategy to create representations of a concept and a mention and compute cosine similarity between the representations of each pair: present a medical code as a single document by concatenating synonymous terms. Then, we apply the TF-IDF transformation on the code and the entity mention and compute the cosine similarity.

Neural networks require word representations as inputs. We investigate the use of several different pre-trained word embeddings. Recent advances have made *distributed word representations* into a method of choice for modern NLP [24, 25, 26]. We utilize word embeddings named *HealthVec*, which are publicly available 200-dimensional embeddings that were trained on 2,607,505 unlabeled user comments (93,526 terms) from health information websites using the CBOW model in [14]. We also experimented with another published 200-dimensional embeddings named *PubMedVec* (2,351,706 terms) trained on biomedical literature indexed in PubMed [27].

4. Experimental Evaluation

The purpose of our evaluation is to determine how well recurrent neural networks can identify the corresponding medical concepts based on informal language from patients' texts.

4.1. Data Set

We conducted experiments on a collection of user reviews obtained from the CADEC corpus [2]. This corpus contains 1,250 reviews and consists of four predefined disease-related types: ADR (6,318 entities), Disease (283 entities), Symptom (275 entities), and Clinical Finding (435 entities). Authors reported that only 39.4% of the annotations (including drugs) were unique; people generally discussed similar reactions. Disease and Symptom specify the reason for taking the drug. Patients may mention the name of a disease or the symptoms that led to them taking a drug. Findings are any adverse side effects, diseases, or symptoms that were not directly experienced by the reporting patient. We did not distinguish between these types and join them into one class of annotations named *Disease*.

All entities in the CADEC corpus were mapped to SNOMED CT-AU (SCT-AU) by a clinical terminologist. SNOMED CT is a clinical terminology that provides codes, synonyms, and definitions of clinical terms, and can be accessed through the UMLS Metathesaurus. Additionally, concepts identified in the SNOMED CT were associated with MedDRA identifiers. In this work, we adopted only SNOMED CT identifiers and removed 'concept less' or ambiguous mentions for evaluation purpose. Table 2 shows final statistics for the CADEC corpus. The total number of unique codes was 1,029.

Table 2: Statistics of the dataset used in the experiments

Entity type	Total	Unique phrases	Unique SNOMED codes
ADR	5,838	3,241	788
Disease	266	165	108
Drug	1,657	290	124
Finding	399	270	180
Symptom	251	128	78

4.2. Preprocessing and Experiment Settings

Preprocessing includes spelling correction and lemmatization using the Natural Language Toolkit (NLTK). We performed a 5-fold cross-validation to evaluate the methods. We found that a standard cross-validation method creates a high overlap of expressions in an exact matching between training and testing parts. Therefore, the split procedure has a specific feature in our setup. First, we removed all duplicates in each dataset. Second, we grouped medical records into sets which are related to a specific medical code. Every such set was split independently into k folds, and all these folds were merged into final k folds. The created folds are publicly available[1].

4.3. Baseline System

For comparison, we applied state-of-the-art baselines based on convolutional neural networks. In [6], experiments showed that CNN outperformed existing strong baselines such as DNorm and Logistic Regression. In order to obtain local features from a text with CNNs, we used multiple filters of different lengths [28].

4.4. Model Configuration and Training

Since neural networks, especially deep neural networks, have a very large number of free parameters, problems with overfitting are inevitable, and some form of regularization is required. We used a dropout rate [29] of 0.5 after the embedding layer (before networks' layers).

Another standard technique in modern deep learning, batch normalisation [30], was designed to cope with a problem known as covariate shift. For all networks, we set the mini-batch size to 128 to minimize the negative log-likelihood of correct predictions.

The last important set of advances deal with actually training the model. We used a popular adaptive gradient descent variations, Adam [31]. Embedding layers are trainable for all networks. The number of outputs of the layer with the softmax activation equals to the number of unique concept codes. Additionally, we separated out 10% of the training set to form the validation set which was used to evaluate different model parameters. The number of epochs is determined by early stopping on the

[1] https://yadi.sk/d/oLBTUpXg3RtCzd

validation set. We employed early stopping after two epochs with no improvement on the validation set. The final number of epochs is 15.

For RNN, we utilized either a 100- or 200-dimensional hidden layer for each RNN chain. For CNN, we adopted effective parameters from [28, 6]. We used the filter w with the window size h of [3, 4, 5], each of which had 100 feature maps. Pooled features were fed to a fully connected feed-forward neural network (with dimension 100) to make an inference, using rectified linear units as output activation.

We found 91% and 88% of words from the CADEC corpus vocabulary in the word embeddings HealthVec and PubMedVec, respectively. For other words, their representations were uniformly sampled from the range of embedding weights [32].

4.5. Results

The standard technique for evaluating concept normalisation is to compare correctly normalised disorder mentions against the gold standard entities [7, 33]. Accuracy which is defined as follows:

$$Accuracy = \frac{N_{correct}}{T_g}, \quad (1)$$

where $N_{correct}$ is the number of correctly normalised disorder mentions and T_g is the total number of disorder mentions in the gold standard.We present the experimental results of neural networks in Table 3. The attention-based GRU with UMLS-based features achieved an accuracy of 69.92%.

Table 3: The accuracy performance of neural networks

Model	Parameters	Accuracy
CNN	HealthVec, 100 feature maps	46.19
CNN	PubMedVec, 100 feature maps	45.79
LSTM	HealthVec, 200 hidden units	64.51
LSTM	PubMedVec, 200 hidden units	64.24
GRU	HealthVec, 200 hidden units	63.05
GRU	PubMedVec, 200 hidden units	62.73
LSTM+Attention	HealthVec, 200 hidden units	65.73
LSTM+Attention	HealthVec, 100 hidden units	64.83
GRU+Attention	HealthVec, 200 hidden units	67.08
with semantic similarity features		
LSTM+Attention	HealthVec, 100 units, similarity TF-IDF	67.63
LSTM+Attention	HealthVec, 200 units, similarity TF-IDF	66.83
GRU+Attention	HealthVec, 100 units, similarity TF-IDF (ALL)	69.92
GRU+Attention	HealthVec, 200 units, similarity TF-IDF (ALL)	69.42

The best results were obtained while using vectors trained on social media posts. GRU consistently outperformed CNNs and LSTM in terms of accuracy. Attention mechanism and prior knowledge from the UMLS Metathesaurus indeed led to quality improvements for both GRU and LSTM.

5. Conclusion

In this work, we have demonstrated that RNN-based architectures, LSTM- and GRU-based in particular, have promising performance on the task of medical concept normalization of free text mentions in social media. The experiments have shown qualitative and quantitative improvement over a strong baseline. We see three possible ways to next research to improve and expand the achieved results. The natural way to extend our models is to integrate a linguistic knowledge into them. We plan to concatenate RNN's output with a semantic similarity vector. We might focus on the development of extrinsic test sets for medical concept normalization. This future work looks promising also in consideration of paraphrase generation and other encoder-decoder applicable tasks.

Acknowledgements

This work was supported by the Russian Science Foundation grant no. 18-11-00284. The authors thank Valentin Malykh for useful discussions during writing this paper.

References

1. *Pradhan S., Elhadad N., Chapman W. W., Manandhar S., Savova G.* (2014), Semeval-2014 task 7: Analysis of clinical text, SemEval COLING, pp. 54–62.
2. *Karimi S., Metke-Jimenez A., Kemp M., Wang C.* (2015), Cadec: A corpus of adverse drug event annotations, Journal of biomedical informatics, Vol. 55, pp. 73–81.
3. *Miftakhutdinov Z., Tutubalina E.* (2017), Kfu at clef ehealth 2017 task 1: Icd-10 coding of english death certificates with recurrent neural networks. CLEF, 2017.
4. *Han S., Tran T., Rios A., Kavuluru R.* (2015), Team uknlp: Detecting adrs, classifying medication intake messages, and normalizing adr mentions on twitter, CEUR Workshop Proceedings, Vol. 1996, pp. 49–53.
5. *Belousov M., Dixon W., and Nenadic G.* (2017), Using an ensemble of generalised linear and deep learning models in the smm4h 2017 medical concept normalisation task, CEUR Workshop Proceedings, Vol. 1996, pp. 54–58.
6. *Limsopatham N., Collier N.* (2016), Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation, ACL.
7. *Aronson A.* (2001), Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, Proceedings of the AMIA Symposium, p. 17.
8. *Leaman R., Doğan R. I., Lu Z.* (2013), DNorm: disease name normalisation with pairwise learning to rank, Bioinformatics, 29(22), pp. 2909–2917.
9. *Leaman R., Wojtulewicz L., Sullivan R., Skariah A., Yang J., Gonzalez G.* (2010), Towards internet-age pharmacovigilance: Extracting adverse drug reactions from user posts to health-related social networks, Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, BioNLP'10, pp. 117–125.
10. *Nikfarjam A., Sarker A., O'Connor K., Ginn R., Gonzalez G.* (2015), Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features, Journal of the American Medical Informatics Association, page 41.

11. *Oronoz M., Gojenola K., Pérez A., Díaz de Ilarraza A., Casillas A.* (2015), On the creation of a clinical gold standard corpus in spanish: Mining adverse drug reactions, Journal of biomedical informatics, Vol. 56, pp.318–332.
12. *Korkontzelos I., Nikfarjam A., Shardlow M., Sarker A., Ananiadou S., Gonzalez G. H.* (2016), Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts, Journal of biomedical informatics, Vol. 62, pp. 148–158.
13. *Sarker A., Nikfarjam A., Gonzalez G.* (2016), Social media mining shared task workshop, Proc. Pacific Symposium on Biocomputing, pp. 581–592.
14. *Miftahutdinov Z., Tutubalina E., Tropsha A.* Identifying disease-related expressions in reviews using conditional random fields, Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog, Vol. 1, pp. 155–167.
15. *Tutubalina E., Nikolenko S.* (2017), Combination of deep recurrent neural networks and conditional random fields for extracting adverse drug reactions from user reviews, Journal of Healthcare Engineering, 2017.
16. *VanDam C., Kanthawala S., Pratt W., Chai J., Huh J.* (2017), Detecting clinically related content in online patient posts, Journal of Biomedical Informatics.
17. *Sarker A., Gonzalez-Hernandez G.* (2017), Overview of the second social media mining for health (smm4h) shared tasks at amia 2017, CEUR Workshop Proceedings, pp. 43–48.
18. *Goodfellow I., Bengio Y., Courville A.* (2016), Deep Learning, MIT Press.
19. *Bengio Y., Courville A., and Vincent P.* (2013), Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell., 35(8), pp. 1798–1828.
20. *Bengio Y., Simard P., and Frasconi P.* (1994), Learning long-term dependencies with gradient descent is difficult, IEEE transactions on neural networks, 5(2), pp. 157–166.
21. *Greff K., Kumar Srivastava R., Koutník J. R., Steunebrink B., Schmidhuber J.* (2015), LSTM: A search space odyssey, CoRR.
22. *Cho K., Van Merriënboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., Bengio Y.* (2014), Learning phrase representations using RNN encoder-decoder for statistical machine translation, arXiv preprint arXiv:1406.1078.
23. *Schuster M., Paliwal K. K.* (1997), Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, Vol. 45(11), pp. 2673–2681.
24. *Goldberg Y.* (2015), A primer on neural network models for natural language processing, CoRR, 2015.
25. *Rubenstein H., Goodenough J. B.* (1965), Contextual correlates of synonymy, Commun. ACM, 8(10), pp. 627–633.
26. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient estimation of word representations in vector space, CoRR, abs/1301.3781.
27. *Moen S., Salakoski T., Ananiadou S.* (2013), Distributional semantics resources for biomedical text processing.
28. *Kim Y.* (2014), Convolutional neural networks for sentence classification, arXiv preprint arXiv:1408.5882.

29. *Srivastava N., Hinton G. E., Krizhevsky A., Sutskever I., Salakhutdinov R.* (2014), Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, Vol. 15(1), pp. 1929–1958.
30. *Ioffe S., Szegedy C.* (2015), Batch normalisation: Accelerating deep network training by reducing internal covariate shift, International Conference on Machine Learning, pp. 448–456.
31. *Kingma D. P., Ba J.* (2014), Adam: A method for stochastic optimization, CoRR, abs/1412.6980.
32. *He K., Zhang X., Ren S., Sun J.* (2015), Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, Proceedings of the IEEE international conference on computer vision, pp. 1026–1034.
33. *Suominen H., Salanterä S., Velupillai S., Chapman W. W., Savova G., Elhadad N., Pradhan S., South B. R., Mowery D. L., Jones G. J., et al.* (2013), Overview of the share/clef ehealth evaluation lab 2013, International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 212–231.

MACHINE LEARNING CLASSIFICATION OF USER INTERESTS ACROSS LANGUAGES AND SOCIAL NETWORKS

Mikhalkova E. V. (e.v.mikhalkova@utmn.ru),
Ganzherli N. V. (n.v.ganzherli@utmn.ru),
Karyakin Y. E. (y.e.karyakin@utmn.ru),
Grigoryev D. A. (Grigd2013@gmail.com)
Tyumen State University (University of Tyumen), Tyumen, Russia

Being a matter of cognition, user interests should be apt to classification independent of the language of users, social network and the essence of interest itself. To prove it, we built a collection of English and Russian Twitter and Vkontakte community pages manually classified according to the interests of their followers. First, we created a model of Major Interests (MaIs) with the help of expert analysis and then classified the mentioned set of pages using machine learning algorithms (SVM, Neural Network, Naive Bayes, Logistic Regression, Decision Trees, k-Nearest Neighbors) trying different optimization techniques. We take three interest domains that are typical of both English and Russian-speaking communities: football, rock music, vegetarianism. The results of classification show a greater correlation between Russian-Twitter and English-Twitter pages. The Logistic Regression with Bernoulli bag-of-words model proves to be the most effective classification algorithm.

Keywords: interest discovery, social networks, natural language processing, optimization

КЛАССИФИКАЦИЯ ИНТЕРЕСОВ ПОЛЬЗОВАТЕЛЕЙ ПРИ ПОМОЩИ МАШИННОГО ОБУЧЕНИЯ В РАЗНЫХ ЯЗЫКАХ И СОЦИАЛЬНЫХ СЕТЯХ

Михалькова Е. В. (e.v.mikhalkova@utmn.ru),
Ганжерли Н. В. (n.v.ganzherli@utmn.ru),
Карякин Ю. Е. (y.e.karyakin@utmn.ru),
Григорьев Д. А. (Grigd2013@gmail.com)
Тюменский государственный университет, Тюмень, Россия

1. Introduction

Social networks provide people with an opportunity to form social clusters that share interests not only sporadically but on a regular basis (circles of fans of different music, books, kinds of sports, etc.). Every circle communicates these interests creating lots of linguistic data to attract new followers and support interests of the existing ones. Researchers often use these data in content-based user models to classify interests of particular users. As a rule, such models are tested on a corpus of one language downloaded from one social network. However, being a matter of cognition, user interests should be independent of the language in which they are expressed and the network where users communicate them, when we try to process them with different algorithms.

To see if the performance of machine learning algorithms is the same for two different languages and two networks, we tested them in three internationally popular interest domains: football, rock music, vegetarianism. For the present research, we collected three datasets from two different networks: the English (I) and Russian (II) corpora from Twitter and the Russian corpus (III) from Vkontakte.[1] Then, we tried to classify the corpora according to the interests of users with such machine learning instruments as SVM, Neural Network, Naive Bayes etc.[2]

One of the most problematic issues in this classification was to find three classes that have enough data in the two networks and two languages. For example, there are lots of pages devoted to football in Vkontakte and Twitter, in Russian and English. However, there appeared to be few Russian vegetarian pages in Twitter, the same was the case for Russian historical reenactors who are only widely present in Vkontakte. The problem of small scale datasets[3] is nowadays addressed quite often irrespective of the branch of science [Huda et al. 2017] in various fields like dialectology, archeology, biology [Plekhanova et al. 2018; Steyerberg et al. 2001] and, closer to our research, recommender systems [Li et al. 2018]. It follows that if we are striving to build a system of interest classification for a social network, we should, first of all, focus on the inherent properties that show in the small but prominent samples and can serve as a standard in large-scale research.

2. Interest discovery by means of NLP

There exists a variety of content-based models of user interests. These models make use of keywords, interests enlisted in profiles, tags attached to posts etc. Such data serve as the classification basis in works of [Bonhard 2006; Firan, 2007; Dugan

[1] The dataset can be found at https://github.com/evrog/TSAAP.

[2] Within the scope of this research, we focus on the classification of pages to learn about the language of social groups in social networks. We apply machine learning and statistical analysis to observe linguistic data rather than to partition a social network into **clusters** of groups and individuals with similar interests. The latter is a practical task that currently has no universally acknowledged solution.

[3] As well as class imbalance [Sitompul et al. 2018].

2007; Li 2008, Sen 2009, Guy 2010][4]. However, they are often very unreliable and hard to formalize. Interest discovery has now become a separate branch of user modelling.

In regard to social networks, NLP provides several approaches to interest discovery. [Piao 2011] view interests as terms and named entities extracted from a collection of user tweets. In works of [Mccallum 2005; Ramage 2010; Ahmed 2011], interests are viewed as topics distributed across users' tweets. The authors apply variations of Latent Dirichlet Allocation suggested by [Blei 2003] as the main method of topic analysis to scale user messages down to a particular topic. [Wang 2014] describe the User Message Model that is designed particularly for microblogs to reduce data sparseness and topic diversity.

Interests can be represented as concepts in an ontology. The latter often includes named entities. [Bakalov 2009] suggest a hybrid user model that makes use of ontologies to specify user interests. Interests are either extracted as keywords from the content of visited pages or can be manually specified by a user. [Al-Kouz 2012] describe another approach where the system creates a semantic graph of interests based on the "entities" mentioned in tweets. Entities are words denoting real-world phenomena that have an encyclopedic description. For reference, the authors used the currently deprecated knowledge base Freebase[5]. At the same time, a recent study of [Piao 2016] demonstrates that "concept-based representations of user interests using a KB" add efficiency to the model, but then there is no need to add "rich semantic information from a KB to extend the interests of users."

3. Modelling social nature of interests

It appears that interest discovery in social networks is a two-sided problem. First, regarding the number of published posts and comments, although in social networks linguistic content is abundant, it is often very hard to structure. Second, user interests themselves are an arcane matter: some researchers view them as topics, tags, keywords, etc. We will call the interest that attracts users to a page, the Major Interest (MaI). In the present research, we will attempt to classify a number of community pages based on three MaIs: football, rock music, vegetarianism.

3.1. Community pages

In our research, we will focus on community pages, e.g. accounts of public value that represent institutions, authorities, famous people, leaders of social groups,

---

[4] In recommender systems, tags and keywords in profiles define a scope of users that share similar interests. According to [Guy 2009], this process is called *collaborative* filtering. [Pazzani 1999] suggests *demographic* filtering that infers types of users with a common interest based on their age, gender, education etc. mentioned in profiles. With the rise of the social network analysis, many researchers, for example [Groh 2007], attempt to objectivize communities with the help of social graphs (*social* filtering). A more detailed account of these approaches is given by [Burke 2002].

[5] http://www.freebase.com. Before the widespread use of knowledge bases, linguists often used WordNet [Stefani 1999]. More recent approaches like [Shen 2013] use DBpedia.

events, etc. They exist in all networks known to us (Twitter[6], Vkontakte[7], Facebook[8], LiveJournal[9] etc.). Many researchers already use data from such pages together with a user's individual page content but view them as complementary material. Usually, but not necessarily, such accounts have many followers (typically, more than 1,000).

Concerning the content downloaded for analysis, from Vkontakte, we obtained posts, comments to posts, and comments from the so-called "board". As for Twitter, the only content available there is tweets.

3.2. Data survey

Observations show that for an expert it is quite easy to bind a community page to one certain MaI based on user comments and tweets and to find other pages with a similar MaI (the same kind of sports, music style, etc.). Many pages even provide links to other recommended pages. However, on the same page, users can mention a variety of different interest domains especially if they are related hyponymically (a style of music and its substyles), antonymically (a football team vs. its opponent in a championship), pragmatically (a football team and a stadium where it trains). Therefore, to define the basis of classification, i.e. MaIs that are not just microtopics and the pages that are devoted to these MaIs, we conducted an expert-based survey.

First, we downloaded comments from 4,000 random Vkontakte community pages that contained from 22 to 100,523 words. Next, we asked a sociologist and a marketing specialist to find several active communities with common interests, i.e. such community pages where people actively interact about something they share an interest for. The result set included four communities whose MaI is one of the following: 1. rock music, 2. historical reenactment, 3. football, 4. vegetarianism. All these MaIs are international and represented by pages in Russian as well as in English. We chose sample discussions from Vkontakte pages where people talk about things related to the MaIs. For control, a sample with several disparate objects of interest was chosen.

10 experts (certified and employed linguists, sociologists, marketing specialists) gave their opinion on what community manifests itself in every sample. We instructed experts to define if authors in the sample dialogue *are* a community and, if yes, explain why they think so. Thus, the expert answers were formulated freely without the aim of interest attribution. Some of them preferred to just name the community ("vegans", "rockers"); some stated the object of interest ("vegetarianism", "rock music"). If these keywords were mentioned, we assigned 1 point to the answer (a True Positive answer); if no or some other keywords were mentioned ("music addicts" instead of "rockers"), we assigned 0 points. The answers were put in a ranking table (cf.: Table 1 in Appendix). In general, agreement between the experts can be considered

[6] https://twitter.com/

[7] https://vk.com/. One of the most popular Russian social networks.

[8] https://www.facebook.com/

[9] https://www.livejournal.com/

reliable, as Krippendorff's α=0.82 (>0.8). To see which samples relate to the most unanimous decision, we calculated percentage of True Positive answers in every column (percent agreement).

Determining adherence of the authors of comments to communities of football fans, vegetarians, and historical reenactors, the raters showed perfect agreement. Fans of rock music were not as easy to define (only 50% of raters recognized them). The control group also provided a highly reliable result[10] that allows us to state that the raters were not apt to see communities in any text we offer them.

3.3. Feature set

After the data survey, we searched Vkontakte and Twitter for pages that attract fans of 1. rock music, 2. historical reenactment, 3. football, 4. vegetarianism. The search showed that historical reenactment has no Russian accounts in Twitter. Hence, we had to exclude it from the further research. For each class in the three corpora (I. English-Twitter, II. Russian-Twitter, III. Russian-Vkontakte), we managed to find a different number of pages from which we downloaded tweets and comments.

Normalization. We parse Twitter pages with our tweet preprocessing software[11]. It has a special treatment of mentions (they start with "@", e.g. "@WhoopiGoldberg" becomes a two word group "whoopi goldberg") and hashtags (e.g. "ElectrikBLOOM" becomes "electric bloom"). In Vkontakte, we remove URLs, attachments and emoji. All texts are converted to lowercase, symbols and punctuation marks are removed.

Lemmatization. The sets are processed as in Normalization (see above), but before the change of case we lemmatize English texts with NLTK Lemmatizer [Bird 2009] and Russian texts with Pymystem3[12].

The properties of the sets are reflected in Table 1.

[10] We assigned 1 point for this sample if the expert directly expressed doubt in describing the community, e.g. wrote "Don't know", "I doubt this is a community at all", or left the field blank.

[11] "Preprocessing tweet" at https://github.com/evrog/PunFields. Its full description can be found in [Mikhalkova 2018].

[12] https://github.com/nlpub/pymystem3

Table 1. Feature set. F—football, R—rock music, V—vegetarianism, T—Twitter, Vk—Vkontakte, En—English, Ru—Russian. Total No. is given as follows: *tokens* first, then *types* (no duplicates). denotes the mean of the scores. In Twitter, the maximum No. of comments downloaded from a page is 1,000; in Vkontakte, the maximum number of wall posts and comments to posts available for download is 100.

		No. of pages	Total No. of words	Total No. of lemmes	No. of words per comment, tweet: $\bar{x}$, mode	No. of comments per page, $\bar{x}$
Vk Ru	**F**	39	738684, 91486	664972, 76657	18.61, 1	992
	R	109	1212731, 136866	1166159, 87589	24.91, 1	438
	V	127	759066, 103800	717531, 62372	58.16, 6	101
T Ru	**F**	33	334457, 38115	330653, 19130	10.64, 12	924
	R	37	312911, 53721	305206, 31538	10.26, 14	802
	V	32	192643, 45042	188852, 26000	11.66, 14	500
T En	**F**	97	1366312, 33321	1726604, 29766	14.36, 15	971
	R	96	960542, 47507	1328049, 40503	11.69, 9	846
	V	100	1189804, 51769	1616783, 43110	12.41, 8	949

As mentioned above, in their study of user interests, researchers mainly appeal to keywords, topics, named entities etc. However, when we asked experts from the Data survey to analyze what makes them think that a page attracts a certain social group, they also pointed at terminology and special meaning of common words, derivation (i.e. words with the same stems: **veg**an, **veg**etarian, **veg**etarianism, etc.) and unique vocabulary (e.g. **hoolie** for football fans). We tend to think that interests cannot be bound to a certain topic or a set of semantically related topics (e.g. football-sports). They are rather like umbrella terms to a combination that singles out an interest from similar ones. For example, the combination of "game-field-ball" differentiates football from hockey ("game-ice-stick"). To make sure that the machine learning classifier learns enough about these differences, we need a sufficient set of words. In the present study, we experiment with 1,000 most frequent items (word forms and lemmes).

4. Community pages classification

We used several machine learning algorithms to classify community pages that represent one of the MaIs: 1. football, 2. rock music, 3. vegetarianism.

4.1. Interclass classification

Cross-validation. The sets of pages being of different size, we split some of the longer texts into smaller ones to create a collection of 200 texts of different length (the total of 1800 texts). We randomly split each set into 5 equal parts to apply *5*-folds cross-validation. Further, we analyze the average of F1-scores.

Classification algorithms that we chose for the survey are often met in NLP tasks like spam detection, sentiment analysis and the like: Support Vector Machine, Neural Network, Naive Bayes, Logistic Regression, Decision Trees, *k*-Nearest Neighbors. Their structure and implementation in the Python library Scikit-learn [Pedregosa et al. 2011] are described in the documentation of the library.

Optimization parameters. We used the following optimization strategies to compare the performance of classifiers. In SVM, we experimented with four kernel functions: linear, polynomial, Radial Basis Function, sigmoid. In Naive Bayes, we separately tried three well-known algorithms based on Bayes' theorem: Bernoulli, Multinomial, and Gaussian.

The Scikit-Learn implementation of Neural Networks uses a Multi-layer Perceptron algorithm. Unlike Logistic Regression, it learns **non-linear** dependencies with the help of hidden layers. We tested the default model with 1 hidden layer of 100 neurons. We also experimented with two solver functions: "*lbfgs* is an optimizer in the family of quasi-Newton methods" and *adam* is "a stochastic gradient-based optimizer" [Pedregosa et al. 2011].

For all the classifiers, we tested three data models: Bernoulli—absence or presence of a word denoted by 0 and 1 correspondingly; Frequency distribution—presence of a word denoted by its frequency in the training vocabulary denoted by a whole number in the interval [0;+∞); Normalized frequency—presence of a word denoted by normalized frequency in the training vocabulary in the interval [0;1]. The lemmatized texts are analyzed separately from the normalized texts.

We also do not exclude stop-words for the following reason. As we deal with social groups, their **use of some stop-words is significant**. First, in Frequency models, such stop-words as "I", "we" and "they" have different frequency. This frequency can show differences in groups' values. For example, some groups can be more focused on collectiveness and use "we" more often than "I"; some can be more competitive using "they", "their", etc. In Bernoulli models, stop-words are not so significant as they are likely to be present in nearly all the texts. However, in cases of short texts, they can be quite important (for example, if there are only words like "we", "us", "our", the group can belong to a more "collective" type).

In case of some classifiers, Scikit-learn offers more instruments for optimization (penalty parameter C and types of loss function in SVM, activation function in neural networks, etc.). Some of them might have been overseen in the experiment, as their scope creates a huge field for testing, or can subject the classifier to over-tuning. However, with our data openly published, we hope they will be further studied in other projects.

4.2. Results of experiment

Table 2 (Appendix) demonstrates average results of F1-score after a 5-fold cross validation. First of all, it shows that lemmatization slightly increases the performance (by about 3%): the sum of -scores of the lemmatized texts is 262.752 versus 254.186 of the normalized texts.

Second, the Bernoulli model is the most effective one by mode: it has 18 scores of 1.0 when the two other models have only 4 such scores together, and by mean: 0.845 against 0.753 for plain and 0.795 for normalized frequencies.

Third, the best performing algorithm is Logistic Regression with Bernoulli model. The sum of its -scores equals 17.71. The second best score (17.664) belongs to the Neural Network (*lbfgs*) with Bernoulli model which hints at the lack of necessity to complicate a Logistic Regression classifier with a non-linear model. The third place belongs to the Multinomial Bayes with plain frequencies (17.5).

SVM models, even the linear one, were not as successful compared to the Logistic Regression. From this, we can assume that the word combinations that help to differentiate between two classes are more different in their core and have blended, noisy margins.

Concerning normalization of word frequencies, it appears to improve performance of such algorithms as SVM with RBF and sigmoid kernels. Without it, SVM 'Sigmoid' shows the lowest results in the ranking table. However, it can also decrease the result. Surprisingly, it derated the average result of Multinomial Naive Bayes from 0.972 to 0.51.

4.3. Statistical analysis

We will now try to analyze differences in classification of the three datasets according to the MaI, the language of user communication and the network where the texts were posted. For the analysis we will use the -scores from Table 2, Appendix. First, we will normalize the Table excluding classifiers that gave lower results in the either of the two sets: normalized, lemmatized.

For every MaI, the total sum of $\overline{F1}$-scores and sum dependent on the language and network is shown in Table 2. We use sums for the first four columns (Total, Vk Ru, T Ru, T En) as every number in each of these columns characterizes sets of the same size. E.g. there are 24 F-scores in "Football, Vk Ru" and "Rock, Vk Ru": 12 F-scores of Bernoulli models and 12 of Frequency models. Hence, there is no need to average them. However, the other four columns characterize sets of different sizes. For example, the "Football—Vkontakte" set of F-scores has 24 items, whereas "Football—Twitter" has 48 items (2 $\overline{F1}$-scores for every English and Russian text corpus of Football fans). Likewise, "Football—Ru" has 48 scores, and "Football—En" has 24 scores. Therefore, we average data in the last four columns. The size of the set of scores in Total is 24×3=72; in Vk Ru, T Ru, T En, it is 24 each.

Table 2. Comparison of MaI $\overline{F1}$-scores. F—football, R—rock music, V—vegetarianism, T—Twitter, Vk—Vkontakte, En—English, Ru—Russian. denotes the mean of the scores

MaI	Total	Vk Ru	T Ru	T En	Vk, $\bar{x}$	T, $\bar{x}$	Ru, $\bar{x}$	En, $\bar{x}$
Normalized texts								
F	33.976	10.240	11.826	11.910	0.853	0.989	0.919	0.993
R	33.138	10.064	11.334	11.740	0.839	0.961	0.892	0.978
V	32.906	9.8080	11.302	10.796	0.817	0.962	0.880	0.983
Lemmatized texts								
F	34.282	10.430	11.932	11.920	0.869	0.994	0.932	0.993
R	34.776	10.398	11.624	11.754	0.867	0.974	0.918	0.980
V	33.708	10.272	11.622	11.814	0.856	0.977	0.912	0.985

To analyze the significance of differences between sets, we used Mann-Whitney test. It supports the following hypotheses:

1. Lemmatized and non-lemmatized (normalized) sets may come from different distributions (i.e. differences in their results are statistically significant): statistic=5296.5, pvalue=0.243 (>0.05), *two-sided*.
2. Differences in the Twitter-Russian and Twitter-English sets are insignificant: statistic=3407.0, pvalue=0.001 (<0.05), *two-sided*. However, the Vkontakte-Russian set underscores significantly compared to the Twitter-English (statistic=161.0, pvalue=1.0, *greater*) and Twitter-Russian (statistic=703.0, pvalue=0.99, *greater*) sets.
3. Vegetarianism and Rock Music are very likely to score less than Football: statistic=1671.0, pvalue=0.99, *greater*, and statistic=1612.5, pvalue=0.99, *greater* correspondingly.

Also, there appears to be no correlation between the experts' difficulty to classify the rock music fans (see 3.2 Data survey) and the ML classification which was successful enough.

In general, the Vkontakte set seems to actually provide a lower performance than Twitter. We suppose that the difference is caused by more noise which can be due to the normalization software. In Twitter, we have a processor for hashtags and mentions that turns them into clear word forms. In Vkontakte, we simply remove all kinds of attachments.

As for the languages, if we do not take into account the mentioned processor, there seems to be no significant differences between the Russian and English languages. For the both languages, lemmatization is a useful tool.

5. Conclusion

Summing up, with due normalization, languages do not influence the ML classification of interests. However, the social network can be an important factor. Which social network features decrease the performance on Vkontakte sets requires more research. Also, there can be differences due to the interest itself (in our case, vegetarianism and rock music were significantly less supple in classification than football).

Concerning the classifiers, we have assumed, on the grounds that the Logistic Regression has the best score, that interest classification is more focused on the core of a set of features rather than the margins. We also faced the efficiency of the Bernoulli model. I.e. word frequencies are not as important in classification as the absence or presence of characteristic features.

If we consider this experiment in terms of a practical application to classify all social network pages according to "user interests", the data in our research is, of course, much more structured. In a real network, it will be hard to get as many expert-classified pages as there are user interests. However, the findings of this research can be helpful in developing practical tools of their discovery.

References

1. *Ahmed A., Low Y., Aly M., Josifovski V., Smola A. J.* (2011), Scalable distributed inference of dynamic user interests for behavioral targeting, In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 114–122.
2. *Alabandi G. A.* (2017), Combining Deep Learning with Traditional Machine Learning to Improve Classification Accuracy on Small Datasets, M.S. thesis submitted to the Graduate Council of Texas State University.
3. *Al-Kouz A., Albayrak S.* (2012), An interests discovery approach in social networks based on semantically enriched graphs, In: Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on, IEEE, pp 1272–1277.
4. *Bakalov F., König-Ries B., Nauerz A., Welsch M.* (2009), A hybrid approach to identifying user interests in web portals, In: IICS, pp 123–134.
5. *Bird S., Loper E., Klein E.* (2009), Natural Language Processing with Python, O'Reilly Media Inc.
6. *Blei D. M., Ng A. Y., Jordan M. I.* (2003), Latent dirichlet allocation, Journal of machine Learning research, 3 Jan, 993–1022.
7. *Bonhard P., Sasse M. A.* (2006), 'Knowing me, knowing you'—using profiles and social networking to improve recommender systems, BT Technology Journal, 24(3), pp. 84–98.
8. *Burke R.* (2002), Hybrid recommender systems: Survey and experiments, User modeling and user-adapted interaction, 12(4), pp. 331–370.
9. *Dugan C., Muller M., Millen D. R., Geyer W., Brownholtz B., Moore M.* (2007), The dogear game: a social bookmark recommender system, In: Proceedings of the 2007 international ACM conference on Supporting group work, ACM, pp. 387–390.
10. *Firan C. S., Nejdl W., Paiu R.* (2007), The benefit of using tag-based profiles, In: Web Conference, LA-WEB 2007, Latin American, IEEE, pp. 32–41.
11. *Groh G., Ehmig C.* (2007), Recommendations in taste related domains: collaborative filtering vs. social filtering, In: Proceedings of the 2007 international ACM conference on Supporting group work, ACM, pp. 127–136.
12. *Guy I., Zwerdling N., Carmel D., Ronen I., Uziel E., Yogev S., Ofek-Koifman S.* (2009), Personalized recommendation of social software items based on social relations, In: Proceedings of the third ACM conference on Recommender systems, ACM, pp. 53–60.
13. *Guy I., Zwerdling N., Ronen I., Carmel D., Uziel E.* (2010), Social media recommendation based on people and tags, In: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, ACM, pp. 194–201.
14. *Huda R. K., Banka H.* (2017), Efficient feature selection and classification algorithm based on PSO and rough sets, Neural Computing and Applications, 1–17.
15. *Li R., Ye X., Zhou H., Zha H.* (2018), Learning to Recommend via Inverse Optimal Matching, arXiv preprint arXiv:1802.03644.
16. *Li X., Guo L., Zhao Y. E.* (2008), Tag-based social interest discovery, In: Proceedings of the 17th international conference on World Wide Web, ACM, pp. 675–684.

17. *McCallum A., Corrada-Emmanuel A., Wang X.* (2005), Topic and role discovery in social networks, Computer Science Department Faculty Publication Series.
18. *Mikhalkova E., Karyakin Y., Voronov A., Grigoriev D., Leoznov A.* (2018), Pun-Fields at SemEval-2018 Task 3: Detecting Irony by Tools of Humor Analysis, In: Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018).
19. *Pazzani M. J.* (1999), A framework for collaborative, content-based and demographic filtering, Artificial intelligence review, 13 (5–6), pp. 393–408.
20. *Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E.* (2011), Scikit-learn: Machine learning in Python, Journal of Machine Learning Research, 12, pp. 2825–2830.
21. *Piao G., Breslin J. G.* (2016), Interest representation, enrichment, dynamics, and propagation: A study of the synergetic effect of different user modeling dimensions for personalized recommendations on Twitter, Springer International Publishing, Cham, pp. 496–510.
22. *Piao S., Whittle J.* (2011), A feasibility study on extracting twitter users' interests using nlp tools for serendipitous connections, In: Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), IEEE Third International Conference on, IEEE, pp. 910–915.
23. *Plekhanova E., Nuzhdin S. V., Utkin L. V., Samsonova M. G.* (2018), Prediction of deleterious mutations in coding regions of mammals with Transfer learning, Evolutionary Applications.
24. *Ramage D., Dumais S. T., Liebling D. J.* (2010), Characterizing microblogs with topic models, ICWSM 10: 1–1.
25. *Sen S., Vig J., Riedl J.* (2009), Tagommenders: connecting users to items through tags, In: Proceedings of the 18th international conference on World wide web, ACM, pp. 671–680.
26. *Shen W., Wang J., Luo P., Wang M.* (2013), Linking named entities in tweets with knowledge base via user interest modeling, In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 68–76.
27. *Sitompul O. S., Nababan E. B.* (2018), Optimization model of K-Means clustering using artificial neural networks to handle class imbalance problem, In IOP Conference Series: Materials Science and Engineering, Vol. 288, No. 1, p. 012075, IOP Publishing.
28. *Stefani A., Strapparava C.* (1999), Exploiting nlp techniques to build user model for web sites: the use of wordnet in siteif project, In: Proc. 2nd Workshop on Adaptive Systems and Comparison of Interest Classifying Model User Modeling on the WWW.
29. *Steyerberg E. W., Eijkemans M. J., Harrell F. E. Jr, Habbema J. D.F.* (2001), Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets, Medical Decision Making, 21(1), 45–56.
30. *Wang Q., Xu J., Li H.* (2014), User message model: A new approach to scalable user modeling on microblog, In: Asia Information Retrieval Symposium, Springer, pp. 209–220.

ANALYSIS OF COREFERENTIAL EXPRESSIONS IN PAWS (ENGLISH-CZECH-RUSSIAN-POLISH PARALLEL TREEBANK WITH ANAPHORIC RELATIONS)

Nedoluzhko A. (nedoluzko@ufal.mff.cuni.cz)[1],
Novák M. (mnovak@ufal.mff.cuni.cz)[1],
Ogrodniczuk M. (maciej.ogrodniczuk@ipipan.waw.pl)[2]

[1]Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic; [2]Polish Academy of Sciences, Institute of Computer Science, Warsaw, Poland

In this paper, we decribe the coreference annotation on a multi-lingual parallel treebank (PAWS), a portion of Wall Street Journal translated into Czech, Russian and Polish which continues the tradition of multilingual treebanks with coreference annotation. The paper focuses on language-specific differences. We analyse syntactic structures concerning anaphoric relations in the languages under analysis, such as personal and impersonal constructions in polypredicative constructions and pro-drop qualities.

Keywords: parallel corpus, multilingual, coreference, Czech, English, Russian, Polish

АНАЛИЗ КОРЕФЕРЕНТНЫХ СВЯЗЕЙ В КОРПУСЕ PAWS (АНГЛО-ЧЕШСКО-РУССКО-ПОЛЬСКИЙ ПАРАЛЛЕЛЬНЫЙ КОРПУС ЗАВИСИМОСТЕЙ С АНАФОРИЧЕСКОЙ РАЗМЕТКОЙ)

Недолужко А. Ю. (nedoluzko@ufal.mff.cuni.cz)[1],
Новак М. (mnovak@ufal.mff.cuni.cz)[1],
Огродничук М. (maciej.ogrodniczuk@ipipan.waw.pl)[2]

[1]Карлов университет, Прага, Чехия;
[2]Польская академия наук, Варшава, Польша

1. Introduction

In recent years, there appeared a number of multi-lingual parallel corpora annotated with referential relations. One of such corpora is the PAWS treebank, which stands for ***P***arallel ***A***naphoric ***W***all ***S***treet Journal. PAWS is a multi-lingual parallel treebank annotated with coreference relations [Nedoluzhko et al., 2018], it is freely available for non-commercial research and educational purposes[1]. Its current release consists of texts in four languages: English (original) and translations into Czech, Russian and Polish.

The aim of this paper is a contrastive analysis of how coreference relations are expressed in particular languages, based on the data from this treebank. The analysis is approached directly by contrasting aligned coreferential expressions in the languages, as it was already done for various expressions in English and Czech [Novák and Nedoluzhko, 2015] and reflexive possessives in English, Czech and Russian [Nedoluzhko et al., 2016a].

As the proposed treebank currently consists of three Slavic languages, it may serve as a valuable source for linguistic research on this language family. However, the translation factor should be taken into account. We deal with the translations from English into Slavic languages, so the direct calques between closely related Czech, Polish and Russian are not possible. On the other hand, translators treat the texts differently: Some of them stay closer to the texts, others try to primarily transfer the meaning, applying the mechanisms of explicitation and implicitation [Blum-Kulka, 1986]. Taking into account the relatively small dataset, the comparison of the resulting structures does not give statistically valuable results, although it gives a number of interesting observations.

The main feature of PAWS is its manual annotation of coreferential relations in all included languages. As two of the languages (Czech and Polish) extensively use zero subjects, we could miss a lot of valuable information if we annotated coreference only on surface. Therefore, we adopted the style based on the theory of Functional Generative Description [Sgall et al., 1986], first used for Czech in the Prague Dependency Treebank 2.0 [Hajič et al., 2006] and for Czech and English in the Prague Czech-English Dependency Treebank 2.0 [Hajič et al., 2012]. In this style, coreference and other anaphoric relations are annotated on the layer of deep syntax called *tectogrammatical layer* which consists of dependency trees containing both explicitly expressed as well as important elided content words. Presence of elided words makes it possible to represent coreferential relations for dropped pronouns as well as for elided noun phrases in some specific syntactic constructions.

To facilitate the cross-lingual analysis, we equip the treebank with word alignment links between all nodes in all languages under analysis, including the reconstructed zeros. **Figure 1** (at the end of the paper) illustrates the annotation of a sample sentence in all four languages, as visualized by the TrEd tool [Pajas and Štěpánek, 2008]. Every sentence is represented as a dependency tree, with squared nodes representing the expressions elided on surface (cf. **`#Cor`** in the English sentence, **`#PersPron`** in the Czech and Polish sentences, etc.). The solid blue and red arrows correspond to coreferential

[1] It can be downloaded from the Lindat/Clarin repository (http://hdl.handle.net/11234/1-2683).

links, word alignment is marked by dashed lines between the nodes in the trees (for clarity, the figure shows only alignment of coreferential expressions).

2. Related work

Our work relates to all multilingual parallel corpora with linguistic annotation, especially those for Slavic languages. ParaSol: A Parallel Corpus of Slavic and other languages [Waldenfels, 2006] is an aligned corpus of translated and original belletristic texts featuring automatic morphosyntactic annotations. The latest version comprises more than 30 languages. InterCorp [Čermák and Rosen, 2012] is another large multi-lingual parallel synchronic corpus with Czech as a pivot language, i.e. every text has its Czech version. It features part-of-speech tagging and lemmatization. The Polish-Russian Parallel Corpus [Laziński and Kuratczyk, 2016] features morphosyntactic description yet both sides differ as far as disambiguation is concerned (present in Polish, absent in Russian part). Paralela [Pezik, 2016] is a translation-based Polish-English corpus based on publicly available multilingual text collections and open-source parallel corpora featuring morphosyntactic annotation.

PAWS is also one of a few corpora annotated with coreference relations. Its English and Czech part directly corresponds to a subset of the Prague Czech-English Dependency Treebank 2.0 [$1] and its coreferential extension [Nedoluzhko et al., 2016b, PCEDT 2.0 Coref] and the Russian part corresponds to the PCEDT-R corpus [Nedoluzhko et al., 2016a], where the texts had been translated into Russian and aligned to Czech and English but they had not been annotated with coreferential relations there. ParCor 1.0 [$1] also belongs to this category. It is a German-English parallel corpus consisting of more than 8,000 sentences. Unlike PAWS, which has annotation of full corerence chains, only pronominal coreference is annotated in ParCor. On the other hand, texts in the corpus come from different genres, which is not the case in PAWS.

3. Data and Basic Statistics

The English texts originally come from the Wall Street Journal section of the Penn Treebank PTB. Czech, Russian and Polish texts have been translated by native speakers of the corresponding languages. English texts with their Czech translations have been extracted from Prague Czech-English Dependency Treebank 2.0 [Hajič et al. 2012]. The data consist of documents located in the first half of the PCEDT section 19 (wsj_1900 to wsj_1949). The basic statistics is shown in **Table 1**.

Table 1: Basic statistics for PAWS

	English	Czech	Russian	Polish
Documents	50			
Sentences	1,078			
Tokens	26,149	25,697	25,704	25,763

All texts have been annotated with rich linguistic information on dependency trees. For Czech and English, the annotation was copied from the PCEDT without any change. For Russian and Polish, the final tectogrammatical trees are slightly simplified and not always guaranteed to be correct, especially as concerns obligatory valency positions of predicates, semantic roles and some types of ellipses.[2]

4. Annotation of Coreference in PAWS

The coreference relations in PAWS have been annotated manually according to the Prague coreference annotation style [Nedoluzhko et al., 2016b]. The annotation covers the cases of grammatical (syntactic) and textual coreference.

The grammatical coreference typically occurs within a single sentence: These are the cases of relative and reflexive pronouns, verbs of control etc. By textual coreference, arguments are not realized by grammatical means alone, but also via context. Within this type, pronominal coreference of personal, possessive and demonstrative pronouns is annotated, as well as coreference with textual ellipsis, nominal textual coreference in case when the anaphoric expression is a full nominal group, anaphoric reference of local and temporal adverbs (*there, then* etc.) and textual reference to multiple antecedents (so-called *split antecedent*).

In case when an anaphoric expression refers endophorically to a discourse segment of more than one sentence, including the cases where the antecedent is understood by inference from a broader co-text, the special relation (*reference to a segment*) is annotated. This kind of relation has no explicitly marked antecedent.

We also have a specifically marked link for *exophora*, which denotes that the referent is "out" of the co-text, i.e. it is only known from the actual situation. Exophoric reference is annotated in case of temporal and local deixis (*this year, this country*), deixis with pronominal adverbs (*here*), as well as exophoric reference to the whole text.

Table 2 shows the statistics of coreference-related annotation in PAWS.

Table 2: Coreference-related annotation in PAWS

	English	Czech	Russian	Polish
Tectogrammatical nodes	18,611	20,696	18,874	18,541
Coreferring nodes	4,210	4,403	4,254	3,371
grammatical coreference	729	528	749	294
textual pron. coref. expressed	544	213	493	206
textual pron. coref. elided	76	643	32	243
textual nominal coreference	1,361	1,496	1,610	1,568
first mentions	1,277	1,330	1,243	979
reference to split antecedents	149	149	91	65
reference to a segment	28	23	16	12
exophora	46	21	20	4

[2] See [Nedoluzhko et al. 2018] for more details.

5. Contrastive analysis of coreference relations statistics

The brief inspection of **Table 2** shows that there are significant differences in the numbers of relations between the languages under analysis. Some of these differences may be caused by the simplification of the tectogrammatical annotation for Polish, and partly also for Russian. For example, we observe that the number of coreferring nodes in Polish is smaller than in the three remaining languages. The reason is that we did not reconstruct all unexpressed valency positions for Polish (e.g. we didn't insert elided Addressee for the verbs of speech (such as *say*, *claim*, *contend*, etc.) which may be connected by coreference relations. Such relations are rather formal, but technically they are missing in Polish, thus reducing the total of coreferring nodes.

Other differences may reflect the varieties in the grammatical structures or different grammatical tendencies in the languages.

For example, in **Table 2**, we observe that the number of tectogrammatical nodes in Czech is larger than in the three remaining languages. This could be caused by the translator's style, in this case it would be the tendency of the Czech translator to larger explicitation [Blum-Kulka, 1986]. However, the manual analysis of the texts shows a strong tendency of Czech to use finite subordinated clauses instead of non-finite infinitive or gerundial clauses in English, Polish and Russian. Finite constructions are natually longer than infinite ones, so the larger number of tectogrammatiacl nodes in Czech could be also explained by this reason. Consider **Example 1**, where, the gerundial clause in English (*continuing a rebound from steep year-ago losses*) is naturally translated into infinite clauses in Polish and Russian, but it is transferred to a finite subordinate clause (*čímž pokračuje v zotavení z velkých loňských ztrát*) in Czech. Both in Polish and Russian, the translation with a finite subordinate clause is also possible, but, as the data show, this is not often the case: On the one hand, infinite constructions are fully acceptable in these two languages, on the other hand, gerundial constructions in English naturally trigger the similar ones in the target language. As for Czech, an infinite clause is not acceptable in this case.

Example 1:

EN: *Morrison Knudsen Corp. posted third-quarter net income of $7.9 million,* ***continuing*** *a rebound from steep year-ago losses.*

PL: *Morrison Knudsen Corp. zaksięgował dochód netto za trzeci kwartał równy 7,9 milionom dolarów,* ***kontynuując*** *odbicie po znacznych zeszłorocznych stratach.*

CZ: *Společnost Morrison Knudsen Corp. vykázala čistý zisk za třetí čtvrtletí ve výši 7.9 miliónu dolarů,* ***čímž pokračuje*** *v zotavení z velkých loňských ztrát.*

RU: *Корпорация Morrison Knudsen опубликовала данные о чистых доходах, составивших $7,9 млн или 69 центов за акцию, в третьем квартале,* ***продолжая*** *восстанавливаться после больших прошлогодних убытков.*

The prevailing personal subordinate clauses in polypredicative constructions with (both expressed and unexpressed) pronouns in Czech also correlates with the

biggest number of coreferring nodes in Czech, as follows from the statistics of the PAWS coreference-related annotation in **Table 2**.

The tendency to impersonal constructions in Polish and Russian is very strong. In some cases, they even tend to be grammaticalized, as in **Example 2**, where the impersonal gerundial constructions *based / bazując / исходя из* function more like secondary prepositions[3]. In this example, the grammatical coreference of the first argument of the gerundial form is problematic, and both in Polish and Russian the use of a gerundial form conflicts grammatical rules of these languages, saying that, e.g. for Russian, an animate subject should be the prototypical coreferential antecedent for the gerund. This conflict is one of the arguments of grammaticalization.

Example 2:

EN: ***Based*** *on the number of Mesa shares* […]*, the proposed takeover would have a value of about $15.3 million.*

PL: ***Bazując*** *na pozostałej liczbie akcji Mesy* […] *proponowane przejęcie osiągnęłoby wartość około 15,3 milionów dolarów.*

RU: *Запланированное поглощение,* ***исходя из*** *количества акций Mesa* […] *имело бы стоимость почти $15,3 млн.*

Another interesting fact following from the coreference annotation statistics in Table 2 is the highest number of grammatical coreference relations in Russian[4], which can be partially explained by a large number of infinitive constructions, where unexpressed subjects are controlled by the actants of their governing control verbs by means of grammatical coreference. In **Example 3**, the infinitive clause *to employ a financial consultant to advise them* is translated with an infinitive clause into Russian, as a deverbative construction into Polish and as a subordinate clause into Czech:

Example 3:

EN: *In response to the specific offer, Gary Risley, Mesa vice president, said management will ask directors* ***to employ*** *a financial consultant to advise them.*

PL: *W odpowiedzi na szczegółową ofertę, Gary Risley, zastępca prezesa Mesy, powiedział, że zarząd poprosi dyrektorów* ***o zatrudnienie*** *konsultanta finansowego w celach doradczych.*

CZ: *Gary Risley, vicepresident společnosti Mesa, uvedl, že jako odpověď na konkrétní nabídku požádá vedení společnosti představenstvo,* ***aby použilo služeb finančního poradce.***

---

[3] In the given example, the gerundial forms in Polish (*bazując*) and Russian (*исходя из*) are very close to the English one (based). However, the syntactic construction is slightly different, so it should not be considered as a calcue.

[4] In Polish, on the contrary, it is very small. The reason for the small number in Polish is the missing annotation of the control verbs coreference.

RU: *В ответ на конкретное предложение Гэри Рисли, вице-президент Mesa, сказал, что руководство попросит директоров* ***нанять*** *финансового советника для получения консультации.*

Interestingly, in **Example 3**, the infinitive construction is the only possible one in the Russian translation. In Polish, the deverbative construction (*o zatrudnienie*) can be changed to the infinitive one or to a finite subordinate clause. In Czech, the finite subordinate clause (*aby použilo služeb finančního poradce*) can be changed to either an infinitive or a deverbative clause.

The difference in corresponding numbers of coreferential nodes in **Table 2** is also influenced by the frequent use of deverbatives in translations in all three Slavic languages. See **Example 4**, where the original finite clause is translated to deverbative clauses into Polish, Czech and Russian.[5]

Example 4:

EN: *Last week, Mesa rejected a general proposal from StatesWest* ***that the two carriers combine****.*

CZ: *Minulý týden společnost Mesa odmítla základní nabídku společnosti StatesWest* ***na sloučení*** *obou přepravců.*

PL: *W zeszłym tygodniu Mesa odrzuciła ogólną propozycję StatesWest* ***dotyczącą połączenia*** *obu przewoźników.*

RU: *На прошлой неделе Mesa отклонила общее предложение от StatesWest* ***об объединении*** *двух перевозчиков.*

Finally, the point of explicitly expressed textual pronominal coreference is especially interesting, as it shows the different degree of pro-drop quialities of English, Czech, Polish and Russian. As observed from **Table 2**, explicitly expressed textual pronominal coreference is most frequent in English (544 cases). Indeed, in English, there is no possibility for subject omission, whereas for Slavic languages this often happens. However, the subject can be omitted in the analysed languages to a different degree. Czech is a highly pro-drop language, where anaphoric use of personal pronouns in the subject position is untypical. On the other hand, Polish and Russian show substantially lower degree of pro-drop qualities, Polish being less pro-drop than Czech, but significantly more pro-drop than Russian. Our numbers here correspond to the analysis in [Kibrik, 2011], where the distribution of pro-drop qualities in these languages is the same. The big number of elided coreferential nodes in Czech (643 relations) also supports this statement.

[5] In this case, this is rather a technical issue pointing on the fact that coreference annotation of the arguments of deverbatives is a very complicated task which was not completed consistently for none of the languages under analysis.

6. Translation factor

The comparison of the parallel sentences in the languages under analysis shows that in many cases the choice of a language expression is not given by the grammatical structure of the corresponding language, but it is triggered by the syntactic structure of the original English sentence. This factor is very important when analysing translated texts and it may potentially explain many statistical differencies. For example, Table 2 gives evidence that coreference is more frequently realized by nominal groups in Russian than in the other languages (1,610 cases). This could be a translation effect that should be however proved by comparison with other translations. The same is true about the difference in the number of tectogrammatical nodes between the languages.

Moreover, the specifity of the texts (mostly business-focused news) causes a number of calcues which make the analysis on the textual level rather problematic.

7. Conclusion

In this work, we presented the basic statistics of coreference-related annotation in the PAWS treebank, a multi-lingual parallel treebank with manual annotation of coreferential relations in English, Czech, Russian and Polish. We proposed explanations to some differences between the languages under analysis, as concerns the number of tectogrammatical nodes, coreferring expressions, grammatical coreference or pronouns. The basic reasons for these differences are (i) in the preferrable use of finite constructions in Czech and infinite constructions in English, Russian and Polish; (ii) in the different pro-drop qualities of the languages. Furthermore, the translation factor is crusial, especially given the relatively small number of the annotated sentences.

8. Acknowledgements

The authors gratefully acknowledge support from the Grant Agency of the Czech Republic (project GA16-05394S) and the Polish National Science Centre (contract number 2014/15/B/HS2/03435). The research reported in the present contribution has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References

1. *Blum-Kulka, Sh.* (1986). Shifts of Cohesion and Coherence in Translation. J. House, Sh. Blum-Kulka (eds): Interlingual and Intercultural Communication. Tübingen: Narr, 17-35.
2. *Čermák, F. and Rosen, A.* (2012). The case of InterCorp, a multilingual parallel corpus. International Journal of Corpus Linguistics, 17(3):411–427.
3. *Guillou, L., Hardmeier, C., Smith, A., Tiedemann, J., and Webber, B.* (2014). ParCor 1.0: A Parallel Pronoun Coreference Corpus to Support Statistical MT.

In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland. European Language Resources Association.

4. *Hajič J., Panevová J., Hajičová E., Sgall P., Pajas P., Štěpánek J., Havelka J., Mikulová M., Žabokrtský Z., Ševčíková-Razímová M., Urešová Z.* (2006): Prague Dependency Treebank 2.0. Software prototype, Linguistic Data Consortium, Philadelphia, PA, USA.
5. *Hajič J., Hajičová E., Panevová J., Sgall P., Bojar O., Cinková S., Fučíková E., Mikulová M., Pajas P., Popelka J., Semecký J., Šindlerová J., Štěpánek J., Toman J., Urešová Z., Žabokrtský Z.* (2012): Announcing Prague Czech-English Dependency Treebank 2.0. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), European Language Resources Association, İstanbul, Turkey, pp. 3153-3160.
6. *Kibrik, A.* (2011). Reference in Discourse. Oxford, United Kingdom.
7. *Laziński, M. and Kuratczyk, M.* (2016). The University of Warsaw Polish-Russian Parallel Corpus. In Ewa Gruszczynska et al., editors, Polish-Language Parallel Corpora, pages 83–95. Instytut Lingwistyki Stosowanej UW, Warsaw.
8. *Nedoluzhko, A., Khoroshkina, A. S., and Novák, M.* (2016a). Possessives in Parallel English-Czech-Russian Texts. Computational Linguistics and Intellectual Technologies, (15): pp. 483–497.
9. *Nedoluzhko, A., Novak, M., Cinková, S., Mikulová, M., and Mírovský, J.* (2016b). Coreference in Prague Czech-English Dependency Treebank. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), pages 169–176, Paris, France. European Language Resources Association.
10. *Novak, M. and Nedoluzhko, A.* (2015). Correspondences between Czech and English Coreferential Expressions. Discours: Revue de linguistique, psycholinguistique et informatique., 16:1–41.
11. *Nedoluzhko, A., Novak, M., Ogrodniczuk, M.* (2018). PAWS: A Multi-lingual Parallel Treebank with Anaphoric Relations. In M. Poesio, editor, CRAC: Computational Models of Reference, Anaphora, and Coreference, co-located with NAACL 2018. USA, New Orleans, The Association for Computational Linguistics.
12. *Pajas, P. and Stěpánek, J.* (2008). Recent Advances in a Feature-rich Framework for Treebank Annotation. In Proceedings of the 22nd International Conference on Computational Linguistics—Volume 1, Stroudsburg, PA, USA. Association for Computational Linguistics.
13. *Pezik, P.* (2016). Exploring Phraseological Equivalence with Paralela. In Ewa Gruszczynska et al., editors, Polish-Language Parallel Corpora, pages 67–81. Instytut Lingwistyki Stosowanej UW, Warsaw.
14. *Sgall, P., Hajicová, E., and Panevová, J.* (1986). The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. D. Reidel Publishing Company, Dordrecht, Netherlands.
15. *Waldenfels, v. R.* (2006). Compiling a parallel corpus of Slavic languages. In B. Brehmer, et al., editors, Text strategies, tools and the question of lemmatization in alignment. Beitrage der Europäischen Slavistischen Linguistik (POLYSLAV 9), pages 123–138. München.

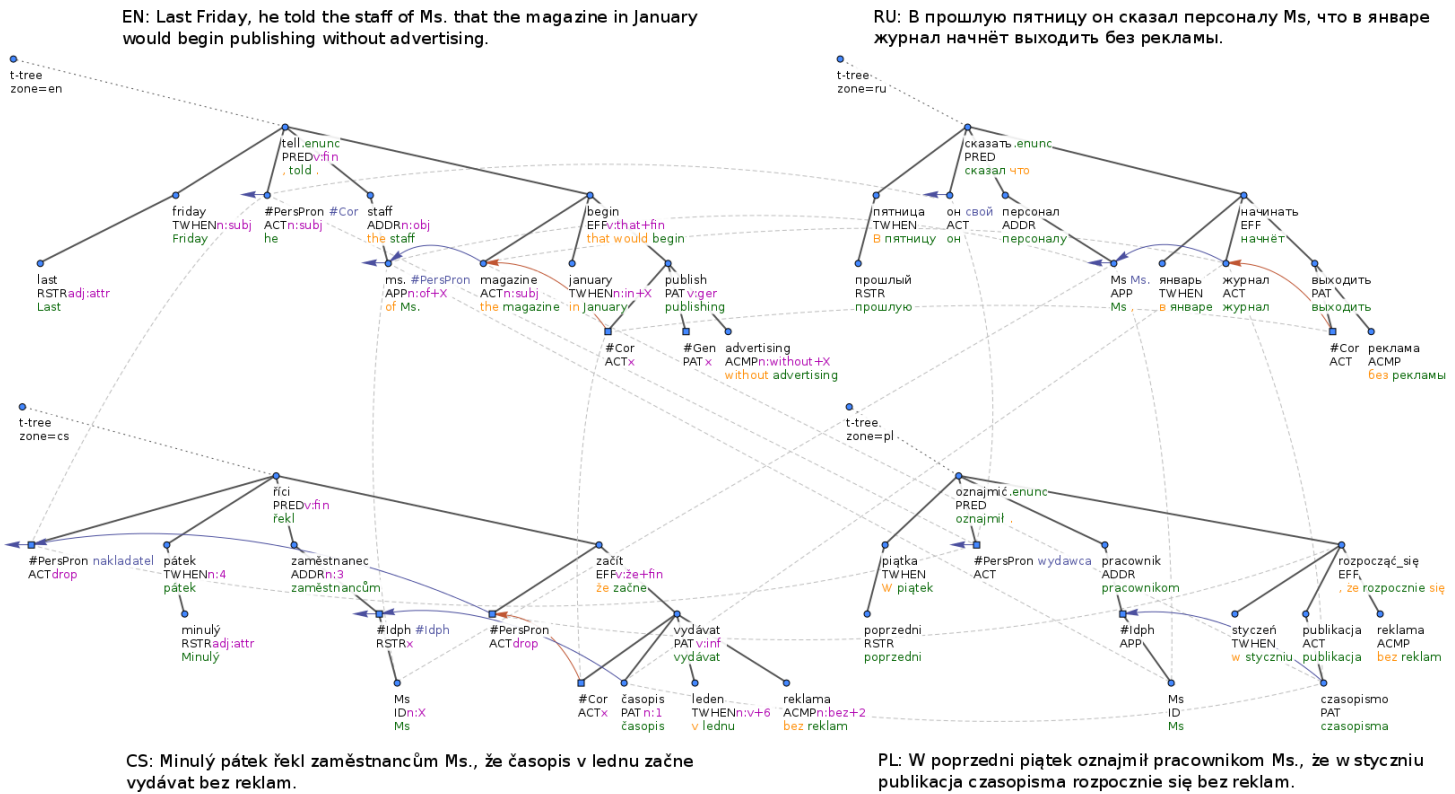


Figure 1: Tectogrammatical representation of a sample sentence in all four languages, visualized by the TrEd tool

PRONOMINAL ADVERBS IN GERMAN AND THEIR EQUIVALENTS IN ENGLISH, CZECH AND RUSSIAN: EVIDENCE FROM THE PARALLEL CORPUS

Nedoluzhko A. (nedoluzko@ufal.mff.cuni.cz)
Charles University, Czech Republic

Lapshinova-Koltunski E. (e.lapshinova@mx.uni-saarland.de)
Saarland University, Germany

The paper presents a contrastive analysis of pronominal adverbs in German (*dabei, darauf, damit* etc.) and their equivalents in English, Czech and Russian. The analysis is based on an empirical study of parallel news texts. Our main focus is to show the interplay between cohesive devices expressed through German pronominal adverbs in text and explore their equivalents in English, Czech and Russian. As the dataset at hand contains translations, we also focus on the influence of the translation factor in parallel texts.

Key words: parallel data, multilingual, translations, comparative study, English, German, Czech, Russian

МЕСТОИМЕННЫЕ НАРЕЧИЯ В НЕМЕЦКОМ И ИХ ЭКВИВАЛЕНТЫ В АНГЛИЙСКОМ, ЧЕШСКОМ И РУССКОМ ЯЗЫКАХ

Недолужко А. Ю. (nedoluzko@ufal.mff.cuni.cz)
Карлов университет, Чехия

Лапшинова-Колтунски Е. (e.lapshinova@mx.uni-saarland.de)
Саарбрюкенский университет, Германия

1. Introduction

The present contribution aims to provide a cross-linguistic analysis of the pronominal adverbs in German (*dabei, darauf, damit* etc.) and their equivalents in English, Czech and Russian. These constructions contribute to the overall textual coherence that is achieved through various types of cohesive devices in a text. These

devices exist in all languages (see e. g. the complex descriptions in de Beaugrande & Dressler (1981), Halliday & Hasan (1976)), but their linguistic realizations depend on the different preferences that languages have. It implies that the distribution of cohesive types may be different across languages, e. g. relations that are typically expressed by connectives in one language may be realized with anaphoric reference in another.

We analyze German pronominal adverbs as they represent multifunctional cohesive devices. For instance, the **German** (DE) pronominal adverb *dabei* (which is a fusion of the preposition *bei* 'at' and the demonstrative pronoun in Dative *dem*)[1] in **Example (1)** may function at the same time as (i) a referring expression, i. e. *dabei* refers anaphorically to the prepositional phrase *beim Betrügen* 'while cheating', and as (ii) a discourse connective expressing a temporal meaning. Moreover, another possible reading of *dabei* in this example is the meaning of contrast and concession. **English** (EN) does not have a direct equivalent for this form. The corresponding example from our parallel dataset contains neither connective, nor anaphoric reference; the discourse relation between the sentences is implicit. The **Czech** translation (CZ) contains the connective *ale* 'but' which has the meaning of contrast and in this case also concession. In **Russian** (RU), the sentence is slightly reformulated and two cohesive devices are used: a conjunction *no* 'but' connecting the clauses in the second sentence and a discourse anaphora *eto* 'it' referring to the event (*обманывать* 'cheat') expressed in the previous clause and in the preceding sentence.

(1) EN (source): *We've learned that a lot of people can cheat. They cheat just by a little bit.*
DE: *Wir haben gelernt, daß viele Leute betrügen können. Der Einzelne betrügt dabei nur ein bißchen.*
CZ: *Zjistili jsme, že hodně lidí je ochotno podvádět. Podvádějí ale pouze po troškách.*
RU: *…люди обманывают. Они обманывают лишь немного, но это всё же обман.*

We are interested in discovering various means that correspond to German pronominal adverbs in English, Czech and Russian. Being very frequent in German, pronominal adverbs (such as *thereby, thereafter, therewith*, etc.) sound rather archaic and are generally avoided in English. In translations from German into English, either multiword expressions consisting of a corresponding preposition and pronoun is used, or this element is dropped. In Slavic languages, prepositional phrases with pronouns are most frequently used, being occasionally lexicalized to the form of pronominal adverbs, similar to the German ones, e. g. Czech and Russian *zato* 'on the other hand'.[2]

Our pilot comparison of the German sentences with pronominal adverbs and their equivalents in English, Czech and Russian suggests the following: (i) Pronominal

[1] There is no clear account on building pronominal adverbs in the grammar studies. In some cases, they are considered to be a fusion of a preposition and pronoun, and a fusion of a preposition and an adverb in the others (see Negele, 2012; DUDEN-Grammatik, 2009).

[2] *zato* 'on the other hand' = preposition *za* 'for' + pronoun *to* 'this'

adverbs are frequent in German and rarely occur in the other languages under analysis; (ii) German pronominal adverbs are ambiguous in their meaning, and therefore, we expect a great variation in their equivalents in the corresponding languages under analysis.

Following these observations, we analyse the usage of the German pronominal adverbs and their equivalents in a multilingual dataset. Our main aim is **to analyse the variation** in the equivalents, and **to describe their functions and usage preferences, trying to find systematicity in their usage**.

The usage of such equivalents is constrained by several factors. First of all, these include existing asymmetries in the language systems. Besides that, at least in the observed data (that contain translations), translation process is expected to have an impact on the choice of such equivalents. For instance, in **Example (1)**, the Czech translator decided to add the contrastive marker *ale* 'but' to explicate the implicit contrastive meaning of the clause, while the German translator prefers to use the pronominal adverb *dabei* which is ambiguous and has both contrastive and temporal readings. We do not know if the usage of this pronominal adverb was triggered by the adverbial *just* in the English original sentence (and transferred into German as *dabei nur* to express contrastive meaning), or *just* corresponds to *nur*, so *dabei* is coreferential, and it was inserted by the translator to create a link between the two sentences for a stylistic purpose. Since the information on the translation process is missing, we are not able to find out translator's motivation.

In this study, we restrict ourselves to the analysis of possible signals in the English sources that trigger the usage of German pronominal adverbs in translations along with their equivalents in the corresponding translations in Czech and Russian. Our observations show that the usage of these constructions in translations may be induced by different constraints. We attempt to explain these constraints with the notion of explicitation borrowed from translation studies.[3]

We believe that our findings will be useful in both theoretical and computational perspectives. The information on the cross-lingual distribution of cohesive means in parallel texts provides the background knowledge for the improvement of multilingual tools for computational discourse analysis. Besides that, the area of machine translation may profit from the information on discourse-aware translation patterns. The knowledge of preferences in the choice of cohesive devices is also important for contrastive linguistics, language learning and translation studies. Moreover, this kind of comparative analysis also provides typologically relevant information on discourse-related phenomena and beyond for each language under analyses.

2. Related Work

Our interest lies on pronominal adverbs classified by Negele (2012:18–20) into conjunctional and phorish (or deictic) ones. To our knowledge, there exist just a few studies that compare several cohesive devices cross-lingually. The only studies known to us include Kunz et al. (2017) on English and German and

[3] The basis of this notion lies in the explicitation hypothesis formulated by Blum-Kulka (1986).

Lapshinova et al. (2015) on English, German and Czech. If we consider empirical studies on German pronominal adverbs, we find just a few example-based ones that address these structures or analyse some cases of their usage. For instance, Dipper & Zinsmeister (2012) mention pronominal adverbs as an interesting task within their coreference analysis of German. Stede & Grishina (2016) consider the anaphoric connective *demzufolge* within the study of discourse relations. Further studies on coreference or discourse connectives in German (e. g. Hinrichs et al., 2005; Krasavina & Chiarcos, 2007; Kunz, 2010) ignore pronominal adverbs, although they constitute around 8% of all referring expressions in German (Lapshinova-Koltunski et al. 2018) and are especially frequent in spoken and spoken-like language. The reason for such a modest attention to this topic in coreference-oriented research can be the fact that pronominal adverbs often (more than 90% in our data) refer to events, whereas coreference research focuses mostly on the entity coreference.

Comparative grammars describe equivalents of German pronominal adverbs in other languages, however, never concerning more than two languages. Some examples of such grammars include Konig & Gast (2012) for English and German, Štícha (2003) for German and Czech, Filippova (2012) for German and Russian, Nelubin (2012) for English and Russian and some others. Another constraint of comparative grammars is that although delivering important knowledge on language contrasts, they are in most cases descriptive and hypothetical. Besides, comparative grammars do not take into account all possible contexts of language use (e. g. spoken vs. written register, formal vs. informal, etc.). Therefore, there is a need in empirical analysis of such phenomena on the basis of multilingual corpora, which is aimed in this paper.

Explicitation (and also implicitation) phenomena have been analysed in a number of corpus-based analyses. However, most of them focus on connectives, e. g. addition (or omission) of (causal) connectives in translations (see Zufferey & Cartoni, 2014, Liu, 2008 or Degand, 2004). For our needs, we adopt Klaudy's definition of explicitation who claims that there are several types of this phenomenon—obligatory, optional, pragmatic and translation-inheritant (see Klaudy, 2008: 106–107). Obligatory and optional explicitation seem to explain the cases that we observe in our data: (1) **obligatory explicitation** occurs due to existing language contrasts—language-specific constructions do not have direct equivalents in a target language, and an element (in our case a pronominal adverb or its equivalent) is added in translation because the target sentence would be ungrammatical without it; (2) **optional explicitation**—the languages under analysis reveal registerial or stylistic differences, the explicitation is optional in the sense that grammatically correct sentences can be constructed also without it, but the text (or a sentence) as a whole will be clumsy and unnatural.

3. Data and Methods

Our approach is data-driven, as we use a parallel corpus to automatically collect the relevant data. Then, the data is manually analysed for transformation patterns that

reflect language differences and the impact of translation process (explicitation/implicitation). The findings are then further interpreted from the point of view of theories, expanding in this way the existing knowledge on the phenomena under analysis.

Unfortunately, there are no corpus resources known to us that would contain German original texts and their translations into English, Czech and Russian. Compilation of such a resource is time-consuming and costly. For this reason, we decide to take an advantage of existing parallel resources, i. e. the translations from English into German, Czech and Russian.

The analysis described in the present paper is based on the preliminary observations that we performed on a different data type, parallel TED talks containing English originals and their translations into German and Czech (Lapshinova et al. 2017). For this paper, we extract news data from the test sets of the translation shared task at the Second Conference on Machine Translation (WMT17, Bojar et al., 2017). We selected 37 English original texts along with their translations into German, Czech and Russian. Both datasets are provided with sentence alignment, so that we do need additional steps to pre-process the corpus. As our primary interest is in German pronominal adverbs, we extract the parallel sentences with these adverbs only. First, we compile a list of such adverbs (*daran, darauf daraus, dabei, dadurch, dafür, dagegen, dahinter, darin*, etc.) using a grammar of German (Duden Online Wörterbuch). Then, we randomly extract 100 corresponding parallel sentences where the aligned German sentence contains one of the pronominal adverbs from the list. After that, we perform a manual alignment of the discourse phenomena in the parallel segments, e. g. connecting *dabei* with corresponding cohesive devices in the other languages, such as *ale* and *no* 'but' in **Example (1)** above. The procedure of manual alignment is one of most important steps of analysis, as it reveals many tiny distinctions and combinations of meanings. The created dataset is then analysed following the questions: What discourse phenomena do pronominal adverbs represent? How are they represented in the source and translations? What are the most frequent transformation patterns and what are the reasons for these particular realizations and transformations?

4. Analysis

4.1. Observations on functions of pronominal adverbs

Pronominal adverbs in German have multiple cohesive functions, as they may refer to either entities or events, or serve as a cohesive conjunction (see **Example (1)** and the clarification in **Section 1** above). In **Example (2)**, the pronominal adverb *damit* may function as a discourse connective expressing causal relations between two propositions. At the same time, it may express an anaphoric reference to the previous context as in **Example (3)**. In this case, *damit* is not a lexicalized connective, but represents a fusion of the preposition *mit* 'with' and the demonstrative pronoun *dem* 'this' referring to the noun *der Gewinn* 'the win'. Pronominal adverbs can be also used in correlative constructions, in which they serve as a sentential proform, as illustrated for an infinitive clause in **Example (4)**.

(2) *Ich besitze keinen Plattenspieler, aber ich würde gerne eine Radiohead Schallplatte kaufen, damit ich sie ins Regal stellen kann* ("*I don't have a record player, but I want to buy a Radiohead record so we can put it on our shelf*").

(3) *Ein brilliantes spätes Tackling von Marcus Watson* [...] *sicherte den Gewinn—und damit die Silbermedaille ("A brilliant late tackle from Marcus Watson* [...] *secured the win—and ultimately the silver medal*").

(4) "[...] *die Bedeutung des Wortes „Patient" habe nichts damit zu tun, Ratschläge zu geben* ("[...] *the word „patient" doesn't mean to make suggestions")*.

In other words, various functions of pronominal adverbs, as well as the necessity of an explicit form in the analysed constructions, depends on a number of factors that may have morpho-syntactic or pragmatic character. The differences between the corresponding devices may also be induced by systemic language differences, i. e. syntactic or morphosyntactic features of one language that do not have direct correspondences in the other languages we are dealing with. For example, some German predicates (verbal, nominal or adjectival) require a prepositional object where English, Czech and Russian predicated do not do so (cf. the German verb *aufhören* 'to stop' in **Example (5)** below requires an explicit preposition object expressed with the pronominal adverb *damit* 'with this', whereas it can be omitted (although presupposed) in English.

(5) EN: *As soon as you win, suddenly stop.*
DE: *Sobald Sie gewonnen haben, hören Sie plötzlich* [***damit***] *auf.*

4.2. Quantitative and qualitative analysis

Table 1 illustrates the quantitative comparison on 100 sentences with pronominal adverbs in German and their equivalents in English, Czech and Russian. The analysis shows that German sentences show more explicitation than their equivalents in Czech and Russian. **In 70% of the observed German translations containing pronominal adverbs, their English sources do not contain any corresponding structure.**

The data also show **a prevalence of correlative uses** (53 vs. 33) for German pronominal adverbs, which gives a significant difference to the observations for the TED talks (Lapshinova et al. 2017)[4]. Correlative use is not triggered by any element in the English source. In Czech and Russian corresponding translations, pronominal proforms are used in 21% and 13% cases respectively. Interestingly, pronominal proforms in both Slavic languages are often optional. A closer look at such cases shows that they are represented by German predicates requiring a prepositional object, i. e. verbs (or nouns and adjectives) whose valency frames contain a preposition. However, this prepositional object is not obligatory in all cases. Therefore, we cannot claim that this is an obligatory explicitation (see **Section 2** above).

[4] In Lapshinova et al. (2017), the relation for 98 sentences was 26 correlative to 63 anaphoric uses, which makes significant difference with $\chi^2 = 18.96$; $df = 2$; $p < 0.001$.

Table 1: Realization of German pronominal adverbs in other languages

Type and # in DE	mapped to	EN abs.	EN in %	CZ abs.	CZ in %	RU abs.	RU in %
anaphoric (33)	zero	17	51.52	18	54.55	17	51.52
	preposition + pronoun	7	21.21	9	27.27	11	33.33
	adverb	1	03.03	2	06.06	1	03.03
	connective	2	06.06	3	09.09	2	06.06
	other (phases, rewordings)	6	18.18	1	03.03	2	06.06
correlative (53)	zero	53	100.00	42	79.25	46	86.79
	preposition + pronoun	0	00.00	11	20.75	7	13.21
connective (4)	connective rewording	4	100.00	3	75.00	3	75.00
other meanings (10)	not analysed						

The qualitative analysis of the data shows that we do observe some cases of obligatory explicitation in our data: The usage of a correlative pronominal adverb or a pronoun in the target languages can be induced by some elements in the source, as illustrated by **Example (6)**. Here, the noun *consensus* is used with a preposition (*on*) in the English source. However, the English language system does not require a proform to add a sentential prepositional phrase, whereas German and Russian do (i. e. the pronominal proforms *darüber* and *в том* 'in that' are obligatory).

(6) EN: *There is no clear consensus on where they can seek common ground on Syria.*
DE: *Es liegt kein eindeutiger Konsens darüber vor, wo ein gemeinsamer Nenner zu Syrien gefunden werden kann.*
RU: *Нет четкого согласия в том, как они могут найти общий язык по Сирии.*

For the **anaphoric function**, we observe that 21% of all occurrences of pronominal adverbs in German were induced by the usage of preposition + pronoun phrases in the English sources. Similarly as for the correlative use, this fact demonstrates the opposite tendency to the one described by Lapshinova et al. (2017) for the TED talks. Ca. 52% of pronominal adverbs do not have any explicit triggers in the English sources (i. e. we observe explicitation in the German translations). The Russian translations seem to reproduce the English sources (no explicitation observed), whereas translations into Czech have even more 'zero' cases—ca. 55% (which maybe an indicator of implicitation). At the same time, Russian and Czech parallel data contain more preposition-pronoun phrases corresponding to the German pronominal adverbs than the English sources do (33% and 27% vs. 21%), which we interpret as an indicator of explicitation in all translations at hand.

The usage of alternative constructions (adverbs, connectives, rewordings) is not considered to be an explicitation, because this is another relation type. However, other interesting observations are possible here. For example, we can find the use of connectives in the source that triggers an anaphoric construction in the German translation.

A closer look at the data reveals the reason—in the English source, there is an elliptical construction *regardless* instead of *regardless of this*, see **Example (7)**. The German translator decides to explicate this ellipsis adding the pronominal adverb *davon* 'of this'.

(7) EN: *But, regardless, Fiji on this form would have beaten a fit as a fiddle 15-man team.*
DE: *Aber unabhängig davon hätte Fiji bei dieser Form ein 15-Mann-Team in Bestform geschlagen.*

Another example of a transformation from a conjunctional into deictic use is illustrated in **Example (8)**. In this case, the German translator prefers not to use a corresponding connective (e. g. *letztendlich*), and uses the pronominal adverb *damit* 'with this' explicating the meaning that the medal was won with what is described in the preceding clause.

(8) EN: *A brilliant late tackle from Marcus Watson… secured the win—and ultimately the silver medal.*
DE: *Ein brilliantes spätes Tackling von Marcus Watson… sicherte den Gewinn—und damit die Silbermedaille.*

If we look at the data from the translatological point of view, we see that Czech and Russian translators keep closer to the English original texts than the German translators do, see **Example (9)**. The German translator added the pronominal adverb *davon* that refers to the nominal phrase *the number of people* in the previous sentence, making the meaning of the interrogative and ambiguous *who* (someone in general or someone out of the stated people who exercise 30 minutes a day) more explicit. The Czech and Russian translators decide to keep this meaning ambiguous.

(9) EN: *Cardiogram… told the Washington Post recently that … the number of people it tracked who did 30 minutes of exercise each day jumped from 45 per cent to 53 per cent. The company does not know who is playing Pokémon Go…*
DE: *Das Unternehmen weiß nicht, wer davon Pokémon Go spielt…*
CZ: *Společnost nemá informace o tom, kdo Pokémon Go hraje…*
RU: *Компания не знает, кто играет в Pokémon Go…*

5. Conclusion

The present paper provided a contrastive study of discourse-related phenomena in four languages on the basis of translation corpora. We selected a set of parallel data consisting of translations into German, Czech and Russian from English, where the German part contains pronominal adverbs that often represent an interplay between different cohesive devices: They are used in multiple functions as anaphoric reference, connectives and correlatives in prepositional sentential clauses. Our study was motivated by the fact that none of the existing studies (both comparative and monolingual grammars) provides information about the contextual and functional preferences of such constructions. However, we were also aware of the specificity of the dataset at hand—translations with only one translation direction—and therefore, we also take into account translation factors influencing such preferences. The

analyses show that German pronominal adverbs are frequently added into translations, when no explicit triggers are present in the English source. Czech and Russian translations seem to be closer to the English sources preserving the source structures. Our qualitative analysis shows that an addition of pronominal adverbs (in both correlative and anaphoric function) is sometimes triggered by the source, e. g. specific constructions in English that are not discourse related (specific phrases, etc.). Besides that, both obligatory and optional explicitation are observed in our data, especially in the German translations. However, we need to treat the result on German with caution, as the data selection was performed on the basis of German pronominal adverbs which are in their nature explicit devices of cohesion.

Nevertheless, we made some interesting observation about genre differences based on our previous analyses. Anaphoric function of pronominal adverbs occurs significantly more frequently in spoken registers. In news, correlative function was predominating, which means that functional preferences of the German pronominal adverbs are context-dependent.

As mentioned above, we are aware of all limitations of our method and data. We know that description of contrastive patterns requires comparative data. At the same time, it is difficult to find multilingual comparative data required for such an analysis (with aligned discourse structures).

Another shortback of our approach is the usage of one translation direction which can, again, be explained by practical reasons—it is difficult to find multilingual translation data with sources and translations available for all the four texts. However, we consider our study to be innovative, as there are no further studies on the interplay between different kinds of discourse phenomena across languages known to us. Besides, we apply a bottom-up approach instead of a top-down one—having corpus data at hand, we analyse different structures trying to understand and explain the observations with the help of existing theories and frameworks. We also make a contribution to these theories and frameworks, as we deliver empirical evidence for the described cases and enhance them with some new phenomena that haven't been yet covered. The results of these analyses are valuable for contrastive linguistics, language learning, translation, and can be applied in the area of multilingual NLP.

Our future work will include a more detailed description of the observed cases, as well as extension of the analysed data. Besides, we plan to have a look at texts translated from German into English, Czech and Russian to be able to make claims about equivalents of pronominal adverbs in these languages. We will also extend our analysis on the two explicitation types and will define a scale for cohesive explicitness as it was done by Zufferey & Cartoni (2014) who defined three degrees of explicitation (no explicitation, light explicitation, strong explicitation). Our scale will be adopted for the cohesive phenomena under analysis that involve not only connectives (as in the study by Zufferey & Cartoni, 2014) but also referential links.

6. Acknowledgements

The authors gratefully acknowledge support from the Grant Agency of the Czech Republic (project GA16-05394S).

References

1. *de Beaugrande, R.-A. and W. Dressler* (1981). Introduction to Text Linguistics. London: Longman.
2. *Becher, V.* (2011). Explicitation and implicitation in translation. A corpus-based study of English—German and German—English translations of business texts. PhD thesis, Universität Hamburg.
3. *Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M.* (2017). Findings of the 2017 conference on machine translation (WMT17). In Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers, p. 169–214, Copenhagen, Denmark, September. Association for Computational Linguistics.
4. *Degand, L.* (2004). Contrastive analyses, translation and speaker involvement: The case of 'puisque' and 'aangezien'. In M. Achard & S. Kemmer (eds.), Language, Culture and Mind, p. 251–270, Stanford: CSLI Publications.
5. *Dipper, S. and H. Zinsmeister* (2012). Annotating abstract anaphora. Language Resources and Evaluation, 46(1):37–52.
6. *DUDEN* (2009). Die Grammatik. Bd. 4.8., überarbeitete Auflage. Mannheim, Leipzig, Wien, Zürich: Dudenverlag.
7. *Duden Online Wörterbuch* (2017) https://www.duden.de/woerterbuch. Accessed 1 December 2017.
8. *Engel, U.* (2004). Deutsche Grammatik. Ein Abriss. München: Hueber.
9. *Filippova I. N.* Sravnitelnaja grammatika Russkogo jazyka. — Moscow, 2012.— 144 pp.
10. *Halliday, M. A.K. & R. Hasan* (1976). Cohesion in English. London, New York: Longman.
11. *Hinrichs, E., S. Kubler, and K. Naumann* (2005). A unified representation for morphological, syntactic, semantic and referential annotations. In Proceedings of the ACL-05 Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, pages 13–20. (TuBa)
12. *Klaudy, K.* (2008). Explicitation. In Baker, M. and G. Saldanha (eds.). Routledge Encyclopedia of Translation Studies, p. 104–108, London-New York: Routledge.
13. *Krasavina, O. and C. Chiarcos* (2007). PoCoS—Potsdam Coreference Scheme. Technical report.
14. *Kunz, K.* (2010). Variation in English and German Nominal Coreference. A Study of Political Essays. Peter Lang.

15. *Kunz, K., Degaetano-Ortlieb, S., Lapshinova-Koltunski, E., Menzel, K. and Steiner, E.* (2017). GECCo—an empirically-based comparison of English-German cohesion. In De Sutter, G. and Delaere, I. and Lefer, M.-A. (eds.). Empirical Translation Studies. New Theoretical and Methodological Traditions. TILSM series. Vol. 300. Mouton de Gruyter, pp. 265–312.
16. *Lapshinova-Koltunski, E., C. Hardmeier and P. Krielke* (2018). ParCorFull: a Parallel Corpus Annotated with Full Coreference. In Proceedings of LREC-2018, Miyazaki, Japan, 7–12 May 2018. ELRA.
17. *Lapshinova-Koltunski E., A. Nedoluzhko and K. Kunz* (2017). Transformations in Discourse: Interplay between DRDs, Coreference and Bridging. Vilnius, Lithuania, ISBN 978-9955-19-883-3, Textlink.
18. *Lapshinova-Koltunski, E., A. Nedoluzhko and K. Kunz* (2015). Across Languages and Genres: Creating a Universal Annotation Scheme for Textual Relation. Proceedings of LAW IX at NAACL HLT 2015. Denver, USA, p. 168–177.
19. *Liu, D.* (2008). Linking adverbials. An across-register corpus study and its implications. International Journal of Corpus Linguistics, 13(4), 491–518.
20. *Meyer, T. and Webber, B.* (2013). Implicitation of discourse connectives in (machine) translation. In Proceedings of the Workshop on Discourse in Machine Translation, pages 19–26, Sofia, Bulgaria. Association for Computational Linguistics.
21. *Nelubin, L.* (2012). Comparative typology of English and Russian. Moscow, Flinta. Nauka.
22. *Negele, M.* (2012). Varianten der Pronominaladverbien im Neuhochdeutschen Grammatische und soziolinguistische Untersuchungen. Studia Linguistica Germanica 108. Berlin, Boston: de Gruyter.
23. *Stede, M. and Y. Grishina* (2016). Anaphoricity in connectives: a case study on German. In Proceedings of the CORBON Workshop. San Diego, California, ACL.
24. *Štícha, F.* (2003). Česko-německá srovnávací gramatika. Praha: Argo.
25. *Weinrich, H.* (2005). Textgrammatik der deutschen Sprache. 3. Revidierte Auflage, Hildesheim, Zürich, New York: Olms.
26. *Zinsmeister, H., Dipper, S., Seiss, M.* (2012). Abstract pronominal anaphors and label nouns in German and English: selected case studies and quantitative investigations. Translation: Computation, Corpora, Cognition, 2(1).
27. *Zufferey S. and B. Cartoni* (2014). A multifactorial analysis of explicitation in translation. Target, 26(3), p. 361–384.

СНЯТАЯ УТВЕРДИТЕЛЬНОСТЬ И НЕВЕРИДИКАТИВНОСТЬ[1]

Падучева Е. В. (elena.paducheva@yandex.ru)
ФИЦ ИУ РАН Москва, Россия

В докладе речь идет о снятой утвердительности (suspended assertion). Показано, что термин снятая утвердительность, который был введен в 1963 году У. Вейнрейхом, охватывает тот же круг явлений, что термин nonveridicality (предлагаемый перевод на русский язык — неверидикативность), который получил широкое распространение в литературе по формальной семантике благодаря работам А. Джаннакиду, Ф. Зварца и др.. Рассматриваются факты русского языка, требующие обращения к понятию снятая утвердительность: местоимения типа *какой-нибудь*, местоимения отрицательной поляризации, исчезновение семантического актанта у глаголов в прямой (не параметрической) диатезе, зеркальная симметрия прошедшего и будущего, отрицание с расширенной сферой действия, местоимения на *-нибудь* в сфере действия отрицания, взаимозаменимость *еще* и *уже*. Высказывается убеждение, что понятие снятой утвердительности будет применяться и в других контекстах.

Ключевые слова: снятая утвердительность, неверидикативность, пропозиция, местоимение

SUSPENDED ASSERTION AND NONVERIDICALITY

Paducheva E. V. (elena.paducheva@yandex.ru)
Informatics and Control Federal Research Centre RAN, Moscow, Russia

The paper addresses the notion of “snyataya utverditel’nost’” (suspended assertion). The author argues that the term “suspended assertion”, introduced by U. Weinreich in 1963, covers the same range of phenomena as the term *nonveridicality* (its suggested Russian equivalent is *neveridicativnost’*), which has become widespread due to the works by F. Zwarz, A. Giannakidou and many others. It is demonstrated that the notion of suspended assertion an be applied to interpret a number of facts of the Russian language,

[1] Работа выполнена при поддержке РФФИ, проект № 17-04-00554.

such as *nibud'*-pronouns, pronouns of negative polarity, the disappearance of a semantic argument of verbs with the direct (non- parametrical) diathesis, the mirror symmetry of past and future, the negation with an extended scope, *nibud'*-pronouns in the scope of negation, the interchangeability of *eshche* 'yet' and *uzhe* 'already'. It's the author's conviction that the notion of suspended assertion will be applicable in many other contexts.

Key words: suspended assertion, nonveridicality, proposition, pronoun

1. Введение

Термин **снятая утвердительность** (suspended assertion) принадлежит У. Вейнрейху (см. [Weinreich 1963], [Veinreikh 1970: 173]). Вейнрейх называет индикатив 'утвердительным наклонением' — в самом деле, индикатив выражает утвердительную, т. е. ассертивную, иллокутивную модальность. Вейнрейх перечисляет средства, которые используются в разных языках для того, чтобы снять утвердительность, изначально присущую индикативу. Это показатели снятой утвердительности (assertion suspending devices), т. е. языковые средства нейтрализации утвердительности (neutralization of assertiveness, или suspension of assertion). К ним принадлежат: конъюнктив, т. е. сослагательное наклонение; императив и вопросительность; отчасти косвенная эвиденциальность и будущее 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, дизъюнкция, целевые и условные союзы; предикаты пропозициональной установки, выражающие неуверенность, предположительность, нереальность.

По [Veinreikh 1970], утвердительная иллокутивная модальность — необходимое условие для того, чтобы пропозиция была соотнесена с реальностью: пропозиция в контексте снятой утвердительности не имеет ни реальной, ни ирреальной объективной модальности, т. е. употребляется **безотносительно к истине**. На самом деле, тут нужна поправка: в контексте фактивного или импликативного глагола / предикатива, когда пропозиция оказывается пресуппозицией или импликативом, она, не имея утвердительной иллокутивной силы, имеет, однако, истинностное значение — является истинной. Например, в контексте (1) пропозиция 'Маша вернулась' истинна—она лишена утвердительной иллокутивной силы, но не истинностного значения:

(1) Я рад, что Маша вернулась.

В любом случае контекст утверждения играет важную роль. В контексте утверждения пропозиция обретает говорящего, который несет ответственность за ее истинность, иначе—берет на себя эпистемическое обязательство (ср. в этой связи парадокс Мура: фраза *Она красива, но я так не считаю* аномальна, поскольку говорящий обязан считать истинным то, что он утверждает).

Ведь и фактивность предиката тоже работает только при условии, что предикат имеет утвердительное употребление. Так, в контексте (2а), где пропозиция с глаголом *уронить* утверждается, отказ представляется говорящим как имеющий место, а в (2б), где этот глагол в сослагательном наклонении и не утверждается, утрачивается и пресуппозиция отказа:

(2) а. Отказ от игры *уронил* Фишера в глазах шахматистов.

б. Отказ от игры *уронил бы* Фишера в глазах шахматистов.

Термин ‘снятая утвердительность’ получил распространение в значении ‘безотносительность к истине’. Он использовался в [Paducheva 1985, с. 94–97; 2004, с. 297–298, 328–329; 2005; 2011; 2013], в [Boguslavskiy (2001, 2008)], [Dobrovol’skiy 2011] и др. Так что за ним можно сохранить право на существование — в следующем значении: пропозиция имеет снятую утвердительность, если говорящий не несет ответственности за ее истинность, т. е. не утверждает ее, а также не предполагает в виде презумпции или следствия.

С конца 90-х годов прошлого века в англоязычной лингвистической литературе возникает термин ‘nonveridicality’ (предлагаемый перевод на русский язык — ‘невериликативность’), см. [Zwarts 1995], [Giannakidou 1998, 1999] и мн. др. Эти работы рассматривают примерно тот круг явлений, которые охватываются термином снятая утвердительность, но в рамках другой теории — в формальной семантике.

Формальная (= теоретико-модельная) семантика сводит смысл к условиям истинности. А термин ‘снятая утвердительность’ возник в рамках семантики, которая определяет смысл с помощью ‘толкований’, т. е. экспликаций (см., например, [Apresyan 1974]; [Wierzbicka 1980]). Эту семантику можно назвать ‘экспликативной’.

Экспликативная семантика оперирует с семантическими представлениями слов и предложений — имея в виду с их помощью объяснить ограничения сочетаемости слов, граммем и конструкций; предсказать массовые сдвиги значения, т. е. регулярную многозначность; моделировать языковые способности говорящего — такие как распознавание синонимии и семантической аномалии (в частности, той, которая лежит в основе понятия пресуппозиции); способность строить высказывание, которое является отрицанием данного, и совершать многие другие операции, предполагающие обращение к смыслу, см. [Apresyan 1999].

Определение невериликативности 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 невериликативным (см. [Zwarts 1995]; [Giannakidou 1998]). Например, контекст *Возможно, Маша вернулась* является невериликативным для пропозиции ‘Маша вернулась’.

Согласно этому определению, невериликативность — это то же, что снятая утвердительность. Как сказано в [Giannakidou 1999], «nonveridical contexts are undefined with respect to truth or falsity». В этих контекстах

пропозиция не определена по отношению к истинности. Контекст отрицания тоже неверидикативный.

Далее я перечислю некоторые языковые сущности или явления русского языка, которые требуют обращения к понятию снятая утвердительность (и неверидикативность):

1) Местоимения типа *какой-нибудь*.
2) Местоимения отрицательной полярности (серия на *-либо* и на *бы то ни было*).
3) Предикаты отрицательной полярности (по [Apresyan V. 2017]): *не нахвалится, не оберешься*.
4) Исчезновение семантических актантов у некоторых глаголов в прямой — не параметрической — диатезе (например, глагола *выбирать*)
5) Зеркальная симметрия прошедшего и будущего.
6) Глобальное отрицание, или отрицание с широкой сферой действия.
7) Местоимение на *-нибудь* в сфере действия отрицания.
8) Взаимозаменимость *еще* и *уже*.

2. Местоимения типа *какой-нибудь*

Вот краткий перечень контекстов, лицензирующих местоимения на *-нибудь*.

1. Отрицание — только в вышестоящей предикации и не во всякой (*Не думаю, чтобы он что-нибудь изменил*).
2. Сопредикатная ИГ с квантором общности (*Каждый что-нибудь принес*).
3. Узуальность и многократность (*Она всегда чем-нибудь недовольна*).
4. а. Условие (*не верь, если кто-нибудь скажет другое*); ограничитель в составе ИГ с универсальной квантификацией — это скрытое условие (*всякий, кто что-нибудь знает о ней, должен это сообщить*); целевой оборот (*спустилась в буфет, чтобы что-нибудь перекусить*).
   б. Обусловленность (*Если Катя дома, в холодильнике уже что-нибудь есть*).
5. Общий вопрос (*Ты ел что-нибудь? Кто-нибудь видел его сегодня?*).
6. Дизъюнкция, т. е. разделительные союзы *или* и *либо…либо* (*Коля или кто-нибудь из его друзей оставил на подоконнике часы*).
7. Модальности возможность и необходимость (*Она могла что-нибудь напутать*; *Надо ей чем-нибудь помочь*).
8. Грамматическое будущее время (*Мы что-нибудь придумаем*).
9. Установки, касающиеся будущего: желание, просьба (*Дай что-нибудь почитать!*), предложение, в том числе — выраженные формой императива; разрешение, согласие, уступка (*хоть*), готовность.
10. Сомнение, предположительность, нереальность и просто мнение; сослагательное наклонение, которое выражает нереальность (*Сомневаюсь, что он что-нибудь сделает*).
11. Сравнение (*Я знаю о вас больше, чем кто-нибудь*).

Это практически весь набор мыслимых контекстов снятой утвердительности, за двумя-тремя исключениями (отрицание, неуверенное восприятие, глаголы речи). Так что снятая утвердительность работает при описании местоимений на *-нибудь*.

3. Местоимения отрицательной полярности

Снятая утвердительность работает при описании местоимений отрицательной полярности. Местоимения отрицательной полярности — это местоименные слова и обороты, которые, не будучи сами по себе отрицательными, не могут употребляться иначе как в контексте отрицания или в других неверидикативных контекстах, в частности в контексте условия или вопроса. В русском языке отрицательную полярность (ОП) имеют местоименные серии на *-либо* (*какой-либо*, *что-либо* и т. д.) и на *бы то ни было* (*какой бы то ни было*, *что бы то ни было* и т. д.). Например, можно сказать: *На Марсе нет каких-либо* (*каких бы то ни было*) *разумных обитателей*; или спросить: *На Марсе есть какие-либо* (*какие бы то ни было*) *разумные обитатели?* Но нельзя сказать: **На Марсе есть какие-либо* (*какие бы то ни было*) *разумные обитатели*. Т.е. местоимения на *бы то ни было* и *либо* невозможны в веридикативном контексте. А возможны в отрицательном и в других неверидикативных.

Языковые единицы с отрицательной полярностью (по-английски — negative polarity items, сокращенно — NPI), не обязательно местоимения, есть в самых разных языках; ср. в английском *any*, *ever*, *yet*, *bother to*, *at all*, *can help* (*doing something*). В русском языке, помимо указанных местоимений, отрицательную полярность имеют слова *настолько, столь*, *больно* в значении ‘особенно’ (как в *не больно надо*) и целый ряд идиом — например, *пальцем* (*не*) *пошевелит*; сочетаний — например, (*не*) *так уж* (*Он мне не так уж нравится / Не скажу, чтобы он мне так уж нравился / Если он тебе так уж нравится…*, но **Он мне так уж нравится*), *такой уж*: *Он не такой уж ловкий*, но **Он такой уж ловкий*.

Из всех контекстов, где возможны местоимения на *-нибудь*, местоимениия ОП допустимы только в контексте 1 — Отрицание (в том числе — прямое и внутрисловное), 4а — Условие, 5 — Вопрос, 11 — Сравнение. Во многих контекстах, где пропозиция употребляется безотносительно к истине, т.е. неверидикативных, местоимения ОП невозможны, например, в контексте модальностей возможность и необходимость, в контексте установок, касающихся будущего (пп. 9 и 10).

3.1. Предикаты отрицательной полярности по Apresyan V. 2017

В [Apresyan V. 2017] приводится обширнейший материал (около 600 единиц по словарю А. И. Федорова и около 100 единиц по МАС), — материал, касающийся отрицательно поляризованных предикатных лексем и фразем русского языка. В основном речь идет не о словах, а о значениях — лексемах: *не врубается, не дождешься*. Различается четыре класса ОП единиц. На семантическом уровне классы различаются по относительному весу ассертивного компонента

и модальной рамки в толковании (модальная рамка — это часть значения, в которой выражается оценка ситуации говорящим и предполагаемое говорящим отношение к этой ситуации со стороны адресата). Я представлю эту классификацию в упрощенном варианте — с точки зрения обязательности отрицания. Класс 1 упрощенной классификации — это класс 1, класс 3 упрощенной — это класс 4, и класс 2 упрощенной — это классы 2 и 3.

Про единицы с отрицательной поляризацией известно, что они обычно допустимы не только в отрицательном контексте, но и в ряде других неверидикативных контекстов, таких как условие или вопрос.

Класс 1 — это единицы с **сильной** отрицательной поляризацией: они не употребляются без отрицания. Примеры лексем класса 1: *не преминул, не взвидел, не нахвалится, не оберешься*. Пример фраземы класса 1 — *не в бровь, а в глаз*. Это самый обширный класс. Он составляет половину всего собрания.

В классе 2 возможно употребление ОП-единицы без эксплицитного отрицания — в контексте снятой утвердительности. Ср.

(1) пальцем не пошевелить — Он прежде чем пальцем пошевелит, будет два часа рассуждать; Если он хоть пальцем пошевелит, я буду удивлена;

(2) не изменится ни на йоту — Вряд ли что-то изменится хоть на йоту;

(3) Нам не впервой ночевать на снегу — Впервой ли нам ночевать на снегу?

(4) Слова не выжмешь — Выжмешь ли хоть слово?

Это единицы второго класса по [Apresyan V. 2017]. Сюда же можно отнести единицы третьего класса по [Apresyan V. 2017] с примерами *Он лыка не вяжет, Я не въезжаю*. Для них тоже возможно употребление без эксплицитного отрицания:

(5) Еле лыко вяжет;

(6) Я с трудом въезжаю.

Правда, не факт, что это контексты снятой утвердительности, т. е. неверидикативные.

Наконец класс 3 упрощенной классификации — это единицы, допустимые не только в контексте снятой утвердительности, но и в утвердительном (веридикативном) контексте. Типичный пример –

(7) Они не афишировали свой роман.

Вполне можно сказать *Они сознательно афишировали свой роман*, т. е. употребить *афишировать* в веридикативном контексте. Но соотношение отрицательных и утвердительных контекстов у *афишировать* по НКРЯ 265: 32.

Сюда же относятся:

(8) Не бери в голову — Но в Москве Егорушка говорил о какой-то Лене, ну я и взяла в голову (Г. Щербакова)

(9) Это ей не по зубам — Вполне по зубам.

Про единицы типа *афишировать, взять в голову, по зубам* можно сказать что, согласно определению, они не являются отрицательно поляризованными; т. е., возможно, являются, но в каком-то другом смысле: по статистике употреблений в отрицательном и в неотрицательном веридикативном контексте.

Так что в статье [Apresyan V. 2017] отрицательная поляризация — это скалярный параметр, каковым она до сих пор не являлась.

4. Исчезновение семантических актантов у некоторых глаголов в прямой — не параметрической — диатезе

Известно, что у глагола СВ *выбрать* есть прямая и косвенная диатеза.

В прямой диатезе, в СВ, есть участник Результат; а в таком же предложении с НСВ в актуально-длительном значении участника Результат нет:

(10) Он выбрал *лошадь* посмирнее [СВ; *лошадь* — конкретно-референтное имя; прямой объект обозначает участника Результат]; [прямая диатеза]

(11) Он выбирает *лошадь* посмирнее [НСВ акт.; *лошадь* не конкретно-референтное имя, обозначает множество выбора; нет Результата выбора]. [косвенная диатеза]

Похожие явления возникают в контексте снятой утвердительности. Так, в (12), где контекст снятой утвердительности создается повелительным наклонением и отрицанием, ситуация не включает участника Результат (поскольку действие относится к будущему в (12а) или не было осуществлено — во всяком случае, не дошло до конца — в (12б)):

(12) а. Выбери *лошадь* посмирнее;

б. Он не выбрал себе *лошади* [прямая диатеза; снятая утвердительность; нет индивидуализированной лошади].

Иными словами, прямой объект глагола СВ *выбрать* обозначает Результат выбора только при утвердительной модальности предложения. В контексте снятой утвердительности прямое дополнение обозначает Множество выбора, как и при НСВ актуально-длительном.

Так что контекст снятой утвердительности оказался нужным при рассмотрении актантной структуры глагола.

5. Зеркальная симметрия прошедшего и будущего

Контекст снятой утвердительности неожиданно проявил себя в аспектологии. Дело в том, что результативное значение, которое усматривается у так наз. общефактического значения несов. вида (*Кто строил* [≈ ‘построил’] *эту дорогу*?), обусловлено ретроспекцией, т. е. порождено прош. временем. В будущем времени, когда нет ретроспекции, значения результативности у формы несов. вида не возникает. Прош. время индикатива порождает у НСВ ретроспективный взгляд на ситуацию и ориентацию на конец, а буд. время дает ориентацию на начало. Это значение не синхронное, но не ретроспективное, как общефактическое, а **проспективное**, см. [Paducheva 2010].

Понятие снятой утвердительности позволяет вскрыть общность в значении повелительного наклонения и буд. времени. В семантике несов. вида императива был в свое время отмечен компонент «внимание на начальной фазе» [см. Paducheva 1996: 68] — непонятного происхождения. Теперь он получил объяснение, когда обнаружился тот же эффект в буд. времени:

(13) а. Звони сейчас = ‘сейчас начинай звонить’ [императив];

б. Сейчас буду звонить матери = ‘сейчас начну звонить’ [будущее время].

В прош. времени активизируется конечная фаза (*Ты звонил матери?*), что порождает импликатуру достигнутого результата; а в будущем активизируется начальная фаза. Поэтому ни в буд. времени, ни в императиве у несов. вида не возникает никаких приращений, свойственных ретроспекции, — ни результативности ни даже прекращения состояния.

Не возникает результативных приращений у несов. вида и в других контекстах снятой утвердительности (а именно — в буд. времени и в ситуации, отнесенной к будущему):

(14) хочу *строить* [≈ ‘хочу начать строить’].

В этой связи объяснился запрет на употребление форм НСВ буд. с глаголами движения:

(15) а. Завтра я буду *строить* гараж, *писать* тезисы на конференцию,

б. *Завтра я буду *ехать* в Крым, *идти* к зубному врачу.

Дело в том, что у глаголов типа *строить* форма НСВ буд. может иметь в том числе и начинательное значение, см. (15а), а у глаголов направленного движения это значение выражается идиоматично — в словообразовании (ср. начинательное значение в *поеду, пойду* и его отсутствие в *построю, напишу*). Формы НСВ буд. времени *буду ехать, буду идти* исключаются по принципу системного вытеснения — как **коровина* из-за *говядина*, при потенциально возможном *ежатина, барсучина*; ср. *баранина*. Синхронное значение возможно:

(16) Завтра в это время я *буду ехать* в Крым;
Дай мне ведро, и я *буду идти* за водой.

Итак, формы буд. и прош. несов. вида зеркально симметричны относительно наст. времени, а не то, что будущее подобно прошедшему.

Эта зеркальная симметричность прош. и буд. проявляется в сочетании с частицей *еще*. В прош. времени *еще* сочетается с СВ только у четырех глаголов (*успеть, остаться, сохраниться, застать*, [Paducheva 2004: 510]), поскольку требует состояния, которое ориентировано на конец, а перфектное состояние у остальных глаголов ориентировано на начало. (В контексте “его еще приняли, а всех остальных прогнали» другое *еще*.)

В буд. времени сочетание с *еще* возможно для всех глаголов СВ:

(17) *еще пришел — еще приду; *еще пожалели — еще пожалеете; *еще поняли — еще поймете.

6. Глобальное отрицание, или отрицание с широкой сферой действия

В [Boguslavskiy 1985] было введено понятие «отрицание с широкой сферой действия» — на примере модификаторов субъектно-предикатного комплекса. Было установлено, что в структуре вида Q (P(x)), где P(x) — предикация, которая представляет собой **субъектно-предикатный комплекс**, а Q — его модификатор, предикатное отрицание в составе P(x), может иметь **широкую** (лучше сказать **— расширенную**) **сферу действия** (включающую Q), если выполняется одно из двух условий:

1) P(x) без Q коммуникативно не значимо;
2) ожидалось, что P произойдет именно в контексте Q.

Так, у предложения (19а), где *Гаврилов прилетит во Львов* — субъектно-предикатный комплекс, а *для участия в конкурсе* — модификатор, есть два разных отрицания, которые оба можно назвать общими: присловное общее отрицание, см. (19б), и предикатное отрицание с расширенной сферой действия, которая включает модификатор, (19в) (термины предикатное и присловное отрицание приблизительно соответствуют англ. sentential negation и constituent negation).

(19) а. Гаврилов прилетит во Львов для участия в конкурсе;

б. (Гаврилов прилетит во Львов) **не** для участия в конкурсе;

в. Гаврилов **не** (прилетит во Львов для участия в конкурсе).

В [Paducheva 2005] было замечено, что, помимо указанных, благоприятным условием для расширенной (иначе — **глобальной)** интерпретации предикатного отрицания является снятая утвердительность. Так, (20б) не является общим отрицанием (20а) и вообще непонятно, что значит. А в (20в) отрицание находится в контексте снятой утвердительности, где оно без труда получает глобальную интерпретацию.

(20) а. Он *резко затормозил*;

б. ?Он *резко не затормозил*;

в. Если бы он *резко не затормозил*, произошла бы авария.

В примерах (21a, b) подчеркнут сентенциальный оператор, который создает для предикации с предикатным отрицанием контекст снятой утвердительности и обеспечивает глобальную интерпретацию этого отрицания. В примере (21b) оператором снятой утвердительности является сослагательное наклонение в определительном предложении, обусловленное несуществованием объекта, обозначенного именной группой.

(21) а. На севере пока *тщательно* не оденешься, из дому не выйдешь;

b. <…> не было такого <человека>, который бы *вскоре* не сделался негодяем.

Сочетания (21a') и (21b'), будучи вынуты из контекста снятой утвердительности, без особой просодии бессмысленны:

(21a') <он> *тщательно* не оделся;

(21b') <он> *вскоре* не сделался негодяем.

Снятая утвердительность отменяет отдельную ассертивность пропозиции P, и структура вида НЕ (P & Q (P)) превращается в структуру вида НЕ Q (P), где Q (P) — единая пропозиция. Q становится как бы дополнительным актантом в пропозиции P.

7. Местоимение на *-нибудь* в сфере действия отрицания

Для того, чтобы построить для предложения с местоимением на *-нибудь* общеотрицательное, надо не только поставить отрицание при глаголе, но и местоимение на *-нибудь* заменить отрицательным местоимением — на *ни-*:

(22) 'НЕВЕРНО, ЧТО (я встретил там *кого-нибудь* из знакомых)' ==>

а. *Я не встретил там *кого-нибудь* из знакомых ==>

b. Я не встретил там *никого* из знакомых.

Т. е. *-нибудь* недопустимо в контексте предикатного отрицания, в сферу действия которого оно входит ([Paducheva 1974: 149], [Pereltsvaig 2000]).

Если местоимение на *-нибудь* таки употреблено в контексте прямого отрицания, оно интерпретируется как не входящее в сферу действия этого отрицания — оно лицензируется каким-то другим оператором снятой утвердительности (в (2) это *если*):

(23) Если *кто-нибудь* не являлся на работу, она воспринимала это как личное оскорбление. ($ > НЕ)

На этом фоне представляют интерес примеры типа (24)–(27). В них местоимение на *-нибудь* находится в сфере действия отрицания (т. е. отрицание не является частным), но не подвергается обязательной замене на отрицательное:

(24) Откусила, подставив снизу ладошку лодочкой, чтобы *чего-нибудь* не уронить. [Ф. Кнорре. Родная кровь (1962)] (НЕ > *нибудь*)

(25) Почти не найти семьи, в которой *кто-нибудь* не пострадал бы [Б. Б. Вахтин. Этот спорный русский опыт (1978)]

(26) Мы спускались осторожно, чтобы *где-нибудь* не сорваться. [Фазиль Искандер. Святое озеро (1969)]

(27) Но по дороге смотрите в оба, чтобы он *куда-нибудь* не свернул! [Н. Леонов, А. Макеев. Ментовская крыша (2004)]

В примерах (24)–(27) замена *-нибудь* на *ни-* не обязательна, но, в принципе, возможна:

(24') Откусила, подставив снизу ладошку лодочкой, чтобы *ничего* не уронить.

(25') Почти не найти семьи, в которой бы *никто* не пострадал.

(26') Мы спускались осторожно, чтобы *нигде* не сорваться.

(27') Но по дороге смотрите в оба, чтобы он *никуда* не свернул!

Такая замена возможна не всегда, но я опускаю сейчас эту подробность.

Эти примеры составляют загадку, поскольку нарушают рассмотренное выше правило о том, что *-нибудь* в сфере действия прямого предикатного отрицания недопустимо.

Очевидно, предложения (24)–(27) демонстрируют специальную конструкцию с нестандартным предикатным отрицанием: отрицание при глаголе совместимо с *-нибудь* в сфере действия этого отрицания и не требует замены местоимения на *-нибудь* отрицательным, Задача состоит в том, чтобы охарактеризовать эту конструкцию в положительных терминах.

А дело в том, что нестандартное отрицание, представленное примерами (10)–(13), есть не что иное как отрицание в контексте снятой утвердительности; оно понимается как глобальное, т. е. включает в свою сферу действия *-нибудь*.

Оператором снятой утвердительности является в примерах (24), (26), (27) *чтобы*, а в примере (25) сослагательное наклонение.

Итак, местоимение на *-нибудь* входит в сферу действия отрицания, поскольку отрицание непосредственно подчинено оператору снятой утвердительности. Этот оператор снятой утвердительности и лицензирует *-нибудь*.

8. Взаимозаменимость *еще* и *уже*

В статье [Boguslavskiy 2002] рассматривается отрывок из поэмы Пушкина «Медный всадник» со странным *уже*:

Еще он думал, что едва ли
С Невы мостов *уже* не сняли <…>

Сначала кажется, что в этом предложении надо *уже* заменить на *еще*: в самом деле, в нем выражается тщетность надежды на то, что мостов еще не сняли.

Однако возможна иная стратегия интерпретации этого фрагмента: *уже не* можно тут понять как *не уже* — точнее, как *неверно, что уже*, так что отрицание будет не частицей, а отдельной предикацией. Как говорит И. М. Богуславский, при определенном условии сентенциальное отрицание *не* может включать *уже* в свою сферу действия — а именно, при условии, что это отрицание входит в контекст квантора, показателя неуверенного мнения (*едва ли, вряд ли, разве, трудно поверить (предположить), сомнительно*); показателя ирреальности, условия и некоторых других. Например:

(28) Трудно себе представить, чтобы к десяти часам вечера они *уже не* выпили все шампанское = ‘трудно себе представить, чтобы было неверно, что к десяти часам вечера они уже выпили все шампанское’.

Круг контекстов, перечисленных И. М. Богуславским, допускает естественное обобщение: все это контексты снятой утвердительности. Правда, в случае *уже* набор этих контекстов заметно уже, чем в случае *-нибудь*, но не в этом суть. Суть в том, что отрицание, попадая в контекст оператора снятой утвердительности (*едва ли*), превращается из частицы, которая, в принципе, имеет ограниченную сферу действия, в сентенциальный оператор со сферой действия более широкой, чем у частицы *уже*. Именно это решение и предлагается в [Boguslavskiy 2002] для *уже* и *не* в контексте снятой утвердительности. Частица *не* ведет себя как оборот *неверно, что*, который составляет отдельную предикацию и занимает семантико-синтаксическую позицию более высокую, чем *уже*. Частица как бы «приклеивается» к оператору снятой утвердительности, сентенциальному, в результате чего сама становится сентенциальным оператором и включает в свою сферу действия *уже*.

Наверняка со временем выяснятся и другие языковые явления, где обращение к снятой утвердительности — она же невериди кативность — будет полезным.

Литература

1. *Apresyan Yu. D.,* (1974). Lexical semantics. [Leksicheskaya semantica] Nauka. Moscow.
2. *Apresyan Yu. D.* (1999). Theoretical semantics at the end of XX century. [Teoreticheskaya semantika v konce XX stoletiya]. Proceedings of the Academy of sciences, Series literature and language, Vol. 58(4), pp. 39–53.
3. *Apresyan V. Yu.* (2017) Negative and positive polarity: semantic sources [Otricatel'naya i polozhitel'naya polyarizaciya: semanticheskiye istochniki] Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2017" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2017"], Vol. 2, pp. 2–15.
4. *Boguslavskij I. M.* (1985) Explorations in syntactic semantics [Issledovanija po sintaksicheskoj semantike] Nauka, Moscow.
5. *Boguslavskij I. M.* (2002). "Sandhi" in syntax ["Sandhi" v sintaksise]. Problems of linguistics [Voprosy jazykoznanija] Vol. 5, pp. 19–37.
6. *Boguslavskij I. M.* (2008). Between truth and lie: adverbials in the context of suspended assertion [Mezhdu istinoj i lozhju: adverbialy v kontekste snjatoj utverditel'nosti. Logical analysis of language. Between lies and fantasy [Mezhdu lozhyu i fantaziey, pp. 67–77. Yazyki russkoy kultury. Moscow.
7. *Dobrovol'skij D. O.* (2011). Comparative phraseology: interlinguistic equivalence and problems of translation of idioms. [Sopostavitel'naja phraseologija: mezhjazykovaja èkvivalentnost' i problemy perevoda idiom] Russian language from a scientific point of view [Russkij jazyk v nauchnom osveshchenii] Vol. 2(22), pp. 219–246.
8. *Giannakidou, A.* (1998). Polarity sensitivity as (non)veridical dependency. Amsterdam, Philadelphia.
9. *Giannakidou, A.* (1999). Affective dependencies. Linguistics and Philosophy, 22(4), 367–421.
10. *Paducheva E. V.* (1974) On semantics of syntax [O semantike sintaksisa] Nauka, Moscow.
11. *Paducheva E. V.* (1985) Utterance and its relation to reality [Vyskazyvanie i ego sootnecennost' s deystvitel'nostyu]. Nauka. Moscow. 6-th edition, 2010, Publishing house LKI, Moscow. http://lexicograph.ruslang.ru/TextPdf1/paducheva1985.pdf
12. *Paducheva E. V.* (1996) Semantic investigations [Semanticheskie issledovaniya] Yazyki russkoy kul'tury, Moscow. http://lexicograph.ruslang.ru/TextPdf1/PaduSemantIssl1996.pdf
13. *Paducheva E. V.* (2004) Dynamic models in the semantics of lexicon [Dinamicheskie modeli v semantike lekciki] Yazyki slavyanskoy kul'tury. Moscow. http://lexicograph.ruslang.ru/TextPdf1/PaduDinamMod2004.pdf
14. *Paducheva E. V.* (2005). Effects of suspended assertion [Ėffekty snyatoy utverditel'nosti: global'noe otricanie] Russian language from a scientific point of view [Russkiy yazyk v nauchnom osveshchenii] Vol/ 2(10), pp. 17–42. http://lexicograph.ruslang.ru/TextPdf2/ryns2005.pdf
15. *Paducheva E. V.* (2010). Mirror symmetry of past and future: the figure of observer. Izvestija RAN [Zerkal'naja simmetrija proshedshego i budushchego: figura

nabljudatelja], Literature and languages series, v.69, №3, pp. 16–20. http://lexicograph.ruslang.ru/TextPdf2/symmetr-2010.pdf
16. *Paducheva E. V.* (2011). Implicit negation and negative polarity pronouns. [Implicitnoe otricanie i mestoimenija s otricatel'noj polyarizaciey] Problems of linguistics [Voprosy yazykoznaniya Vol.1, pp. 3–18. http://lexicograph.ruslang.ru/TextPdf1/vnutrilex_neg-VJa.pdf
17. *Paducheva E. V.* (2013). Russian negative sentence [Russkoe otricatel'noe predlozhenie] Yazyki slavyanskoy kul'tury. Moscow.
18. *Pereltsvaig, A.* (2000). Monotonicity-based vs. veridicality-based approaches to negative polarity: evidence from Russian. In T. H. King & I. A. Sekerina (Eds.), Formal Approaches to Slavic Linguistics (FASL-8).The Philadelphia Meeting 1999 (Michigan Slavic Materials, Vol. 45, pp. 328–346). Ann Arbor.
19. *Veinreikh U.* (1970). On the Semantic Structure of Language [O semanticheskoj structure jazyka] Novoe v zarubezhnoj lingvistike, vol. V. Language universals [Jazykovye universalii] Progress, Moscow, pp. 163–249.
20. *Weinreich U.* (1963) On the Semantic Structure of Language. In: J. Greenberg, ed., Universals of Language, Cambridge, MA: MIT Press, pp. 114–171.
21. *Wierzbicka, A.* (1980). Lingua mentalis. The semantics of natural language. Sydney.
22. *Zwarts, F.* (1995). Nonveridical contexts. Linguistic Analysis, Vol. 25(3–4), pp. 286–312.

RUSSE'2018: A SHARED TASK ON WORD SENSE INDUCTION FOR THE RUSSIAN LANGUAGE[1]

Panchenko A. (panchenko@informatik.uni-hamburg.de)
University of Hamburg, Hamburg, Germany

Lopukhina A. (alopukhina@hse.ru)
National Research University Higher School of Economics, Moscow, Russia
The Vinogradov Institute of the Russian Language, Russian Academy of Sciences, Moscow, Russia

Ustalov D. (dmitry@informatik.uni-mannheim.de)
University of Mannheim, Mannheim, Germany
Krasovskii Institute of Mathematics and Mechanics, Yekaterinburg, Russia

Lopukhin K. (kostia.lopuhin@gmail.com)
Scrapinghub, Moscow, Russia

Arefyev N. (nick.arefyev@gmail.com)
Lomonosov Moscow State University, Moscow, Russia
Samsung Moscow Research Center, Moscow, Russia

Leontyev A. (aleksey_l@abbyy.com)
ABBYY, Moscow, Russia

Loukachevitch N. (louk_nat@mail.ru)
Lomonosov Moscow State University, Moscow, Russia

The paper describes the results of the first shared task on word sense induction (WSI) for the Russian language. While similar shared tasks were conducted in the past for some Romance and Germanic languages, we explore the performance of sense induction and disambiguation methods for a Slavic language that shares many features with other Slavic languages, such as rich morphology and virtually free word order. The participants were asked to group contexts of a given word in accordance with its senses which were not provided beforehand. For instance, given a word "bank" and a set of contexts for this word, e.g. "*bank* is a financial institution that accepts deposits" and "river *bank* is a slope beside a body of water", a participant was asked to cluster such contexts into *unknown in advance* number of clusters corresponding to, in this case, the "company" and the "area" senses of the word "bank". For the purpose of this evaluation campaign, we developed three new evaluation datasets based on sense inventories that have different sense granularity. The contexts in these datasets were sampled from texts of Wikipedia, the academic corpus of Russian, and an explanatory dictionary of Russian. Overall, 18 teams participated in the competition submitting 383 models. Multiple teams managed to substantially outperform competitive state-of-the-art baselines from the previous years based on sense embeddings.

Keywords: lexical semantics, word sense induction, word sense disambiguation, polysemy, homonymy

[1] The authors listed in random order as they contributed equally to this study.

RUSSE'2018: ДОРОЖКА ПО ИЗВЛЕЧЕНИЮ ЗНАЧЕНИЙ СЛОВ ИЗ ТЕКСТОВ РУССКОГО ЯЗЫКА[2]

Панченко А. (panchenko@informatik.uni-hamburg.de)
Гамбургский университет, Гамбург, Германия

Лопухина А. (alopukhina@hse.ru)
НИУ «Высшая школа экономики», Москва, Россия
Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Усталов Д. (dmitry@informatik.uni-mannheim.de)
Университет Мангейма, Мангейм, Германия
Институт математики и механики им. Н. Н. Красовского УрО РАН, Екатеринбург, Россия

Лопухин К. (kostia.lopuhin@gmail.com)
Scrapinghub, Москва, Россия

Арефьев Н. (nick.arefyev@gmail.com)
Московский Государственный Университет им. М. В. Ломоносова, Москва, Россия
Московский Исследовательский Центр Самсунг, Москва, Россия

Леонтьев А. (aleksey_l@abbyy.com)
ABBYY, Москва, Россия

Лукашевич Н. (louk_nat@mail.ru)
Московский государственный университет им. М. В. Ломоносова, Москва, Россия

В статье описываются результаты первого соревнования по автоматическому извлечению значений слов из неразмеченного корпуса текстов для русского языка. Подобные соревнования проводились для некоторых романских и германских языков; мы исследуем методы извлечения значений и разрешения многозначности на материале одного из славянских языков, обладающих богатой морфологией и достаточно свободным порядком слов. Участникам соревнования было предложено сгруппировать контексты слова в соответствии с его значениями, причем сами значения необходимо было автоматически извлечь из корпуса текстов. Например, для неоднозначного слова «замок» нужно было выделить неизвестное заранее число кластеров, соответствующее его значениям, и классифицировать контексты этого слова так, чтобы каждый контекст попал в тот или иной кластер, соответствующий значению слова — «сооружение» и «устройство, препятствующее доступу куда-либо» для контекстов слова «замок». Для оценки качества работы методов мы подготовили три набора данных, различающихся, во-первых, гранулярностью значений и, во-вторых, источниками контекстов (статьи русскоязычной Википедии, материалы Национального корпуса русского языка и толкового словаря). В соревновании приняли участие 18 команд, приславших 383 моделей. Качество результата, полученного представленными моделями, превосходят эталонные методы, основанные на векторах смыслов.

Ключевые слова: лексическая семантика, извлечение смыслов, разрешение лексической многозначности, полисемия, омонимия

[2] Все авторы внесли равный вклад в работу; порядок авторов выбран случайным образом.

1. Introduction

RUSSE[3] is a series of workshops on evaluation of semantic models for the Russian language. The first workshop on semantic relatedness and similarity was held in 2015 in conjunction with the Dialogue conference [Panchenko et al., 2016][4]. The second event, described in this paper, is dedicated to Word Sense Induction (WSI). Word sense induction is the process of automatic identification of word senses in raw corpora. While evaluation of various sense induction and disambiguation approaches was performed in the past for the Western European languages, e.g., English, French, and German, no systematic evaluation of WSI techniques for Slavic languages are available at the moment. This shared task makes a first step towards bridging this gap by setting up an evaluation campaign for one Slavic language. The goal of this campaign is to compare sense induction systems for the Russian language. Many Slavic languages[5] still do not have lexical resources of broad coverage providing a comprehensive inventory of word senses like the English WordNet. Therefore, word sense induction methods investigated in this shared task can be of great value enabling semantic processing of under-resourced languages and domains.

The contribution of our work is two-fold. First, we present the first shared task on word sense induction for a Slavic language. Second, we present three novel sense annotated datasets with about 17 thousand sense-annotated contexts from three sense inventories.

This paper is organised as follows: In **Section 2**, we describe previous shared tasks covering other languages. In **Section 3**, we outline the proposed evaluation methodology. **Section 4** describes three evaluation datasets. **Section 5** presents top scored systems participated in the task. Finally, **Section 6** summarizes key results of the shared task.

2. Related Work

In this section, we start with an overview of shared tasks on word sense induction assuming no sense inventory is provided. All prior shared tasks on this topic were conducted for the English language during the SemEval competitions. Next, we briefly overview previous approaches to word sense disambiguation and induction.

2.1. Shared Tasks on Word Sense Induction

In 2007, SemEval participants were provided with 100 target words (65 verbs and 35 nouns), each target word having a set of contexts where the word appears [Agirre and Soroa, 2007]. A part of these contexts was given as a train set, the rest served as a test set. Average number of senses in the dataset was 3.68 per word. Two

[3] https://russe.nlpub.org

[4] http://www.dialog-21.ru/en

[5] http://sigslav.cs.helsinki.fi

evaluation scenarios were proposed. The first scenario was the evaluation of the induced senses using evaluation metrics for clustering. The obtained clusters were compared to the sets of examples labeled with the given gold standard word senses (classes), and evaluated using the clustering measure called FScore. FScore is calculated as the average of the best F-measure values for each cluster relative to the gold standard classes. The second scenario is the mapping of the induced senses to the gold standard senses and using this mapping to label the test corpus with gold standard labels. The results are evaluated with the precision and recall measures for supervised word sense disambiguation systems. It was found that the FScore measure penalized systems with a high number of clusters, and favored those that induced less senses. Supervised evaluation seemed to be more neutral regarding the number of clusters, as the ranking of systems according to this measure include diverse cluster average. So the ranking of the systems varies according to the used evaluation method.

In 2010, a similar evaluation was devoted to word sense induction for 100 words [Manandhar et al., 2010]: 50 nouns and 50 verbs. For each target word, participants were provided with a train set in order to learn the senses of that word. Then, participating systems disambiguated unseen instances (contexts) of the same words using the learnt senses. The organizers used two other measures of evaluation in comparison to the 2007 task: paired F-score calculated as F-measure of example pairs included or not-included in the induced clusters; and V-measure which assessed the quality of a clustering solution by explicitly measuring its homogeneity and its completeness according to gold standard classes. It was found that V-measure tended to favor systems producing a higher number of clusters. The organizers concluded that the current state-of-the-art lacks unbiased measures that objectively evaluate clustering.

In 2013, the evaluation was focused on the multi-sense labeling task [Jurgens and Klapaftis, 2013]. In this setup, participating systems annotate a context with one or more sense labels weighted by the degree of applicability, which implies the use of fuzzy clustering methods. Measuring the quality of clustering requires handling overlapping clusters, for which two new evaluation measures have been proposed: fuzzy B-Cubed and fuzzy normalized mutual information.

2.2. Word Sense Disambiguation and Induction for Russian

For Russian, Loukachevitch and [Chuiko 2007] studied the all-word disambiguation task on the basis of the RuThes thesaurus. They experimented with various parameters (types of the thesaurus paths, window size, etc). [Kobritsov et al. 2005] developed disambiguation filters to provide semantic annotation for the Russian National Corpus[6]. The semantic annotation was based on the taxonomy of lexical and semantic facets. In [Lyashevskaya and Mitrofanova, 2009], statistical word sense disambiguation methods for several Russian nouns were described.

For word sense disambiguation, word sense frequency information is very important. Loukachevitch and [Chetviorkin 2015] studied the approach of determining the most frequent sense of ambiguous words using unambiguous related words and

[6] http://ruscorpora.ru/en

phrases described in the RuThes thesaurus. [Lopukhina et al., 2018] estimated sense frequency distributions for noun taken from the Active Dictionary of Russian.

Concerning word sense induction task for Russian, [Lopukhin et al. 2017] evaluated four methods: Adaptive Skip-gram, Latent Dirichlet Allocation, clustering of contexts, and clustering of synonyms. [Ustalov et al. 2017] proposed a fuzzy graph clustering algorithm Watset designed for unsupervised acquisition of word senses and grouping them into sets of synonyms (synsets) using semi-structured dictionaries, such as Wiktionary and synonymy dictionaries.

3. Shared Task Description

This shared task is structurally similar to prior WSI tasks for the English language, such as SemEval 2007 WSI [Agirre and Soroa, 2007][7] and SemEval 2010 WSI&D [Manandhar et al, 2010][8] tasks. Namely, we rely on the "lexical sample" setting, where participants are provided with a set of polysemous words, each word is provided with a set text fragments called *contexts* representing examples of the word usage in various senses.

For instance, the contexts for the word "bank" can be "In geography, the word *bank* generally refers to the land alongside a body of water" and "The *bank* offers financial products and services for corporate and institutional clients". For each context, a participant specifies the sense of the target word. Note that we do not provide any sense inventory: the participants can assign sense identifiers of their choice to a context, e.g., "bank#1" or "bank (area)". The only requirement is that the contexts with the different senses of the target word should be assigned with the different identifiers, while the contexts representing the same senses should be assigned with the same identifier. In our study, we use the word "context" as the synonym of the word "instance" used in SemEval [Agirre and Soroa, 2007]; [Manandhar et al, 2010]. Detailed instructions for participant were provide on the shared task[9] website and in the GitHub repository.[10]

3.1. Tracks

We distinguish two tracks in RUSSE'2018. In the *knowledge-free* track, the participants induce a sense inventory from any text corpus of their choice and use this inventory for assigning sense identifiers to the contexts. In the *knowledge-rich* track, the participants use an existing sense inventory, i.e., a dictionary, to disambiguate the target words. The use of the gold standard inventories are prohibited in both tracks.

The advantage of our setting is that virtually any existing word sense induction approach can be used within the framework of our shared task, starting from

[7] http://semeval2.fbk.eu/semeval2.php?location=tasks&taskid=2

[8] https://www.cs.york.ac.uk/semeval2010_WSI

[9] https://russe.nlpub.org/2018/wsi/

[10] https://github.com/nlpub/russe-wsi-kit

unsupervised sense embeddings to the graph-based methods that rely on lexical knowledge bases, such as WordNet.

3.2. Evaluation Datasets

We provide three labeled datasets with contexts sampled from different text sources, which are based on different sense inventories. Each of the dataset was split into the train and test sets. Both sets use the same corpora and annotation principles, but the target words are different. The train set was given to the participants for tuning their models before the competition starts. The test set was made available without labels at the end of the competition. We provide an extensive description of the datasets in **Section 4**.

3.3. Quality Measure

Similarly to SemEval 2010 Task 14 and SemEval 2013 Task 13 on word sense induction and disambiguation, we use a gold standard, in which each polysemous target word is provided with a set of contexts. Each context is manually annotated with a sense identifier as according to the predefined sense inventory. A participating system assigns the sense identifiers from the chosen sense inventory to these ambiguous contexts, which can be seen as clustering of contexts. Thus, to evaluate a system, the labeling of contexts provided by the system is compared to the gold standard labeling, although the sense inventories are different.

Clustering-based measures have an important constraint: they provide contradictory rankings. For instance, none of the five evaluation measures in the SemEval 2013 shared task agree to each other, preferring larger or smaller clusters, see [Jurgens and Klapaftis 2013]. In our shared task, we wanted to avoid having multiple evaluation measures that may provide conflicting results. Moreover, we wanted to have a measure which is equal to zero in the cases of trivial clustering, i.e., random clustering, separate cluster for each context, single cluster for all contexts. We selected a measure that fits all these demands, namely the Adjusted Rand Index (ARI) by [Hubert and Arabie, 1985]. We adopted ARI implementation from the scikit-learn library[11]. The measure was also used before for evaluation of word sense induction in SemEval 2013 Task 11 [Navigli and Vannella, 2013] and in [Bartunov et al., 2016].

3.4. Baseline Systems

We provided a state-of-the-art baseline based on unsupervised word sense embeddings called AdaGram [Bartunov et al., 2016], which is a multi-prototype Bayesian extension of the Skip-gram model [Mikolov et al., 2013]. We rely on a model by [Lopukhin et al. 2017] trained on the 2B tokens-large lemmatized corpus combining the ruWac Internet corpus [Sharoff, 2006], the Russian online library lib.ru, and the Russian Wikipedia. The baseline was obtained using the following hyperparameters:

[11] http://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html

the maximum number of senses of 10, the sense granularity of 0.1, the vector dimension of 300, and the context window of 5. No additional tuning of baseline method on the train data was performed; its performance could be further improved by adjusting the of number of senses for each dataset, merging of similar senses and weighting the contexts. In addition to AdaGram, we provided trivial baselines based on random assignment of word senses, putting each context into a singleton cluster, and putting all the contexts of a word into the same cluster.

4. Evaluation Datasets

We prepared three new gold standard datasets for RUSSE'2018. These datasets are complementary in terms of the average number of senses per word (granularity) of their sense inventories and in terms of the text corpora from which the contexts were sampled. Each of these datasets is named by *corpus-inventory* principle. We have also provided the participants with three published datasets from [Lopukhin and Lopukhina, 2016] as a source of additional training data. Statistics for all the datasets used in the shared task are presented in **Table 1**.

Table 1. The datasets used in the shared task. The "main" datasets were used to test the runs of the participants, and the "additional" datasets were provided as a source of extra training data

Dataset	Type	Inventory	Corpus	Split	# of words	# of senses	Avg. # of senses	# of contexts
wiki-wiki	main	Wikipedia	Wikipedia	train	4	8	2.0	439
wiki-wiki	main	Wikipedia	Wikipedia	test	5	12	2.4	539
bts-rnc	main	BTS	RNC	train	30	96	3.2	3,491
bts-rnc	main	BTS	RNC	test	51	153	3.0	6,556
active-dict	main	Active Dict.	Active Dict.	train	85	312	3.7	2,073
active-dict	main	Active Dict.	Active Dict.	test	168	555	3.3	3,729
active-rnc	addi-tional	Active Dict.	RNC	train	20	71	3.6	1,829
active-rutenten	addi-tional	Active Dict.	ruTenTen[12]	train	21	71	3.4	3,671
bts-rutenten	addi-tional	BTS	ruTenTen	train	11	25	2.3	956

[12] The ruTenTen11 is a large web-based corpus of Russian consisting of 18 billion tokens, which is available thought the Sketch Engine system [Kilgarriff et al., 2004].

4.1. *wiki-wiki:* A Dataset Based on Wikipedia

This sense inventory was built from scratch using words from homonymous word forms dictionary[13] and their senses occurred in the Russian Wikipedia article titles. The contexts have been extracted from the Russian Wikipedia. We assumed that given a Wikipedia article containing an ambiguous word in its title, all the occurrences of this word in this article will share the same sense. Hence, we manually assigned sense identifiers to the titles and extracted contexts of these senses from the full texts of the articles automatically. The datasets contains 9 nouns with 20 homonymous senses.

To construct the dataset, list of the Russian Wikipedia articles which titles contain homonyms from the dictionary has been created. These homonyms which do not occur in the article titles or occurred less than 40 times in the corresponding articles were excluded. The titles for each of the remaining words were grouped manually as according to the homonym sense. Each sense was described using related words (synonyms, antonyms, associations etc.) from the Russian Wiktionary. This resulted in the sense inventory an excerpt of which is presented in Table 2.

Table 2. An excerpt from the sense inventory of wiki-wiki dataset: the word "белка"

word	articles	sense
белка	кавказская белка; обыкновенная белка; японская белка; капская земляная белка; аризонская белка; ...	рыжая, шустрая, дерево, вскарабкаться, спрыгнуть
белка	домен белка; биосинтез белка; фолдинг белка; институт белка ран; сигнальная функция белка; ...	желток, пища, углевод, рацион, жир
белка	белка и стрелка; белка и стрелка (мюзикл); белка и стрелка. лунные приключения; ...	космос, полет, животные, первые, советские

Then, for each sense of each homonym we parsed full texts of the corresponding articles and extracted each occurence of the homonym with at least 50 words to the left and at least 50 words to the right to form a context. If we found no at least 10 contexts for any sense of a homonym, we excluded it with all its senses from the dataset to keep the dataset balanced. Finally, all the contexts have been verified by the organizers; only 9 out of 15 homonyms were left.

4.2. *bts-rnc:* A Dataset Based on the Russian National Corpus

This dataset is based on the sense inventory of the Large Explanatory Dictionary of Russian[14] (*Bolshoj Tolkovyj Slovar'*, BTS; [Kuznetsov, 2014]. The contexts were

[13] http://cfrl.ruslang.ru/homoforms/index.htm

[14] http://gramota.ru/slovari/dic

sampled from the Russian National Corpus (RNC, 230 million tokens in the main corpus)[15]. The train set contains 27 ambiguous words: 6 polysemous words with metaphorical senses and 21 homonymous word[16]. The test set contains 51 ambiguous words: 12 polysemous words with metaphorical senses and 39 homonymous word. We selected these two types of ambiguity—homonymy and metaphorical extension in polysemy—because they were proven to be distinguishable by native speakers in psycholinguistic experiments [Klein and Murphy, 2001]; [Klepousniotou, 2002]; [Klepousniotou et al., 2008]. In this shared task, 29 out of 60 homonyms have only one sense each (e.g. "крона" as a "crown of a tree" and "крона" as in "Norwegian or Danish krone"), the other 31 homonyms were polysemous (e.g. "икра" as "roe / caviar" or "eggplant paste" and "икра" as "calf of a leg"). So we assumed they might be also distinguishable in language models.

The dataset was manually annotated by four students majoring in linguistics. Then, the experts checked the annotation and fixed the mistakes. To ensure the high quality of the annotated dataset, we invited expert linguists for a systematic check of every annotated context for complex words with a high number of polysemous senses. For simpler words with a small number of homonymous senses, we used microtask-based crowdsourcing. Namely, 20 words and 2,547 contexts were checked using crowdsourcing, and 58 words and 7,500 contexts were checked by 7 human experts, which are the authors of this paper. Each human expert read all the contexts and fixed wrong sense annotations, or removed contexts which were too ambiguous or simply irrelevant, e.g., in the cases when a real sense mentioned in the context was actually not in the sense inventory. Overall, 2,103 out of 12,150 contexts were removed, such as an irrelevant context for the word "гвоздика" presented below. This word representing "flower" and "spice" senses are confused in this context with its homograph "гвоздик" (nail):

> … *Посмотри, как здорово это будет выглядеть! — Хорошо, а гвоздики для картин вы сами в стенку вбиваете?* …
>
> … *Look how great it will look! — Well, do you drive nails for the pictures into the wall?*

Another example of the filtered sentences, is with an ambiguous context for the word "крыло" (wing) where the described situation is unclear:

> *… волны, а чуть противный ветер, и крылья повисли; рядом же мчится, несмотря ни на что, пароход, и человек сидит*
>
> … *waves, but a slightly nasty wind, and the wings are hanging down; nevertheless, a steamship is racing alongside and a man is sitting*

[15] http://ruscorpora.ru

[16] In the case of homonymy, a lexical item carries two (or more) distinct unrelated meanings, such as bank as a financial institution and bank as a side of a river; in the case of polysemy senses of a word are related, e.g. blood in "His face was covered in blood" and "They had royal blood in their veins". [Lyons, 1977]

The crowdsourcing annotation was performed on Yandex.Toloka platform[17]. For annotation, we used a subset of words with two or three distinct meanings. In this task, a crowd worker is provided with a set of contexts with a highlighted word to be disambiguated. The worker chooses one of the senses listed below the sentence and submits the answer. The workers demonstrated a high inter-annotator agreement as according to the Krippendorff's α value of 0.825 [Krippendorff, 2013].

4.3. *active-dict*: A Dataset Based on a Dictionary

The Active Dictionary of Russian is an explanatory dictionary that has a strong theoretical basis in sense distinction and reflects contemporary language. (*Aktivnyj slovar' russkogo jazyka*; [Apresjan, 2014]; [Apresjan et al., 2017]). The word senses in the Active Dictionary are considered distinct if they have different semantic and syntactic properties, collocational restrictions, synonyms, and antonyms. For each sense, we extracted all examples (short and common usages) and illustrations (longer, full-sentence examples from the Russian National Corpus) that were used as context in this shared task. On average, we extracted 22.9 contexts per word. The train set, having 85 ambiguous words (84 polysemous words and 1 homonym) and 2,073 contexts, was extracted from publicly available first and second volumes of the dictionary (letters A–G; [Apresjan, 2014]). The test set, having 168 ambiguous words (167 polysemous words and 1 homonym), and 3,729 contexts, was taken from the third volume of the dictionary that became available in March 2018 (letters D–Z; [Apresjan, Galaktionova, Iomdin, 2017]).

To construct the dataset, we extracted examples and illustrations for all polysemous nouns and merged homonymous nouns together. The parser inputs an unstructured representation of the dictionary in a word processor format and outputs a set of labeled contexts.

5. Participating Systems

Overall 18 teams participated in the RUSSE'2018 shared task. We provide here self-descriptions of the approaches used by the teams ranked within the top 5 list in each dataset. The descriptions for all the models submitted by the participants for all datasets can be found in the CodaLab platform in the "Public Submissions" section. We denote each team with its CodaLab login, e.g., "jamsic", and also provide a reference to the paper describing the approach, where available. Note that all the participants submitted to the "knowledge-free" track and we received no submissions to the "knowledge-rich" track.

[17] https://toloka.yandex.ru

5.1. The *wiki-wiki* Dataset based on Wiktionary

17 teams submitted 124 runs for this dataset, with the top teams being:

- **jamsic**. This team used a pre-trained CBOW word embeddings model with 300 dimensions based on the Russian National Corpus by [Kutuzov and Andreev, 2015][18]. The sense clusters were obtained directly from this model by looking at the list of the nearest neighbours. The approach identifies two senses per word. First, the most similar term to a target word is retrieved. This word represents the first sense. Second, vector representation of this word is subtracted from the vector of the target word and again the most similar term is retrieved. This second term represents the second word sense. Disambiguation of a context is performed via calculation of cosine distance of a context representation (an average of embeddings) with these two prototypes. [Pelevina et al. 2016] proposed another method for induction of senses from word embeddings which used clustering of ego-network of related words. However, this approach does not make use of vector subtraction operation employed by the *jamsic* team.
- **akutuzov** [Kutuzov, 2018]. This team used Affinity Propagation to cluster weighted average of word embeddings for each context. The embedding model was trained on the Russian National Corpus using a newer version of the embeddings as compared to.
- **ezhick179** [Arefyev et al., 2018]. This team used Affinity Propagation to cluster the non-weighted average of CBOW vectors for contexts trained on a large corpus of Russian books based on the lib.rus.ec collection with the vector dimensions of 200, the context window of 10, in 3 iterations [Arefyev et al., 2015].[19]
- **aby2s**. This team relied on hierarchical clustering of context embeddings based on the Ward clustering with cophenetic distance criterion and a threshold of 2.6. Sentences were represented as normalized sums of fastText [Bojanowski et al., 2016] embeddings pre-trained on a Wikipedia corpus.
- **Pavel** [Arefyev et al., 2018]. This team used agglomerative clustering of the weighted average of Word2Vec vectors for contexts. The words were weighed using the $\text{tfidf}^{1.5} \times \text{chisq}^{0.5}$ score. The word embeddings were the CBOW vectors for contexts trained on lib.rus.ec with the vector dimensions of 200, the context window of 10, in 3 iterations [Arefyev et al., 2015].

5.2. The "bts-rnc" Dataset based on the Russian National Corpus

16 teams submitted 121 runs for this dataset, with the top teams being:

- **jamsic, akutuzov, ezhick179, Pavel** used methods described in **Section 5.1**.
- **fogside**: Used word embeddings trained on a combination of Wikipedia, Librusec and the training dataset. A neural network with self-attention was used to encode the sentence representations, which were subsequently clustered with the k-means algorithms with k=2.

[18] http://rusvectores.org

[19] https://nlpub.ru/RDT

5.3. The "active-dict" Dataset based on a Dictionary

18 teams submitted 138 runs for this dataset, with the top teams being:

- **jamsic**. This team used a pre-trained CBOW word embeddings model of 300 dimensions based on the Russian National Corpus by [Kutuzov and Andreev, 2015] as in the previous two datasets. However, in this submission the authors followed the approach to word sense embeddings proposed by [Li and Jurafsky, 2015].
- **akutuzov, ezhick179, Pavel**: These teams used methods described in **Section 5.1**.

6. Results

Tables 2, 3, and 4 present the results of the shared task for the three datasets used for evaluation: *wiki-wiki*, *bts-rnc*, and *active-dict*. Each table lists top 10 best teams with the public and private ARI scores on the test set (see **Section 4**). We disregarded from the final ranking teams which were created by organizers for testing purposes[20] and teams which did not provide any description of the used approach[21]. The private ARI scores are used for final ranking of the participants, while the public scores were visible to the participants on the leaderboard immediately after submission before the final deadline. Private and public scores were calculated on non-overlapping sets of words, with public words constituting approximately one third of all words in test set of each dataset. Public scores allowed participants to immediately see their position relative to other participants, while using private scores for final evaluation ensured that participants did not pick their submission based on the leaderboard score. A large difference in public and private scores for the *wiki-wiki* dataset is due to the public set consisting of contexts for just two words: this caused large variance between the public and the private parts for this dataset. However, private/public test differences are substantially smaller for other larger datasets.

Table 3. Top 10 teams out of 17 on the "wiki-wiki" dataset. The full table is available at the CodaLab platform: https://competitions.codalab.org/competitions/17810#results

Rank	Team	ARI (public)	ARI (private)
1	jamsic	1.0000 (1)	0.9625 (1)
2	akutuzov [Kutuzov, 2018]	0.9823 (2)	0.7096 (2)
3	ezhick179 [Arefyev et al., 2018]	1.0000 (1)	0.6586 (3)
-	*akapustin*	0.6520 (6)	0.6459 (4)
-	*aby2s*	1.0000 (1)	0.5889 (5)
-	*bokan*	0.7587 (5)	0.5530 (6)
*	**AdaGram [Bartunov et al, 2016]**	**0.6278 (7)**	**0.5275 (7)**
4	Pavel [Arefyev et al., 2018]	0.9649 (3)	0.4827 (8)
5	eugenys	0.0115 (12)	0.4377 (9)
6	mikhal	1.0000 (1)	0.4109 (10)
7	fogside	0.6520 (6)	0.3958 (11)

[20] Team names: russewsi, lopuhin, panchenko, dustalov.

[21] Team names: joystick, Timon, thebestdeeplearningspecialist, bokan, akapustin, ostruyanskiy.

Table 4. Top 10 teams out of 16 on the "bts-rnc" dataset.
The full table is available at the CodaLab platform:
https://competitions.codalab.org/competitions/17809#results

Rank	Team	ARI (public)	ARI (private)
1	jamsic	0.3508 (1)	0.3384 (1)
2	Pavel [Arefyev et al., 2018]	0.2812 (2)	0.2818 (2)
-	*joystick*	0.2477 (5)	0.2579 (3)
-	*Timon*	0.2360 (7)	0.2434 (4)
3	akutuzov [Kutuzov, 2018]	0.2448 (6)	0.2415 (5)
4	ezhick179 [Arefyev et al., 2018]	0.2599 (4)	0.2284 (6)
-	*thebestdeeplearningspecialist*	0.2178 (8)	0.2227 (7)
5	fogside	0.1661 (10)	0.2154 (8)
*	**AdaGram [Bartunov et al., 2016]**	0.2624 (3)	0.2132 (9)
-	*aby2s*	0.1722 (9)	0.2102 (10)
6	bokan	0.1363 (11)	0.1515 (11)

Table 5. Top 10 teams out of 16 on the "active-dict" dataset.
The full table is available at the CodaLab platform:
https://competitions.codalab.org/competitions/17806#results

Rank	Team	ARI (public)	ARI (private)
1	jamsic	0.2643 (1)	0.2477 (1)
2	Pavel [Arefyev et al., 2018]	0.2361 (4)	0.2270 (2)
-	*Timon*	0.2324 (5)	0.2222 (3)
-	*thebestdeeplearningspecialist*	0.2297 (6)	0.2194 (4)
3	akutuzov [Kutuzov, 2018]	0.2396 (3)	0.2144 (5)
-	*aby2s*	0.2465 (2)	0.1985 (6)
-	*joystick*	0.1890 (8)	0.1939 (7)
4	ezhick179 [Arefyev et al., 2018]	0.1899 (7)	0.1839 (8)
*	**AdaGram [Bartunov et al., 2016]**	0.1764 (9)	0.1538 (9)
-	*ostruyanskiy*	0.1515 (10)	0.1403 (10)
-	*akapustin*	0.1337 (11)	0.1183 (11)

Several observations can be made on the basis of the results presented in the Tables 3–5. First, the method of **jamsic**, based on extraction of sense inventory directly from word sense embeddings showed good results on two datasets were it was applied substantially outperformed all other methods (see **Section 5** for a detailed description of the methods). A particularly large advantage of this method over other method, which relied on some kind of sentence clustering, is observed for the coarse-grained wiki-wiki dataset, because it contains only homonymous senses, which can be easily extracted with such an approach. On the *active-dict* dataset this participant also outperformed other teams, but in this case using the approach of [Li and Jurafsky, 2015] to the construction of word sense embeddings.

Second, other approaches showing good results ranking in top 2–5 were the methods based on direct clustering of textual contexts represented with the features based on word embeddings pre-trained on large corpora, such as the Russian National Corpus or a collection of books from the lib.rus.ec library. In particular, successful methods relied on the Affinity Propagation clustering approach, but also some other methods, such as Agglomerative and Ward clustering algorithms. The **fosgide** team used word embeddings and the *k*-means clustering algorithm. Namely, for each context, a context vector is built as a non-weighted average of the fastText vectors for the words in the context. The context vectors for each target word are decomposed into a linear combination of learnt basis vectors. Then, weight vectors of this decomposition are clustered using k-means clustering algorithm.

The Affinity Propagation method is well-suited in the case of word sense induction task as it defines the number of parameters automatically, in contrast to, e.g., Agglomerative Clustering, which produces a lot of senses in the case of this task, as the number of sense per word is usually distributed according to a power law. Nevertheless, the *Pavel* [Arefyev et al., 2018] team managed to obtain two second-best results on two datasets using Agglomerative Clustering with a fixed number of clusters (different in the case of each dataset, learned from the train data). It was shown that a carefully selected weighting schema for words can provide an edge with respect to a un-weighted average of word embeddings. Besides, on *wiki-wiki* and *rnc-bts* datasets, *jamsic* team provided good results with the method which also yields two senses per word for all words. In case of the first dataset, this could be explained by the fact that the average polysemy of this dataset is 2. In the case of the second dataset with more senses, the good performance could be explained by a skewed distribution of senses across the sentences: the majority of the contexts belong to two senses (which is not the case for the sense-balanced *active-dict* dataset).

Third, multiple teams managed to outperform a competitive baseline provided by the organizers based on the AdaGram [Bartunov et al., 2016] word sense embeddings. There is a substantial difference in ARI score between different datasets. The scores for *wiki-wiki* dataset are much higher due to a low number of senses per word and extremely clear separation between senses. Among two other datasets based on dictionary senses, scores for *bts-rnc* are higher than for *active-dict* due to *active-dict* using a more granular sense inventory and having a much smaller number of contexts per sense (just 6.8 instead of 42 for *bts-rnc*), see Table 1 and analysis in [Lopukhin et al., 2017]. Another difference is that for *bts-rnc* contexts were randomly sampled from corpus, while contexts for *active-dict* were selected by the authors of the Active Dictionary of Russian, with both full sentences from the corpus and short usage examples — it remains unclear how this difference contributed to the difference in scores. Still, ARI scores for all datasets are higher than what was reported in [Bartunov et al. 2016] for SemEval-2007 and SemEval-2010 datasets for word sense induction.

Finally, one can observe a large difference in absolute scores for the coarse-grained *wiki-wiki* dataset and the two datasets based on fine-grained word sense inventories coming from dictionaries (*active-dict* and *bts-rnc*). Discriminating between a large number of related polysemous senses is naturally a more challenging task, which requires more sophisticated representations and methods. We hope that the

setup of our shared task will pave the way towards developing methods which are able also to excel on these more challenging datasets.

7. Conclusion

In this paper, we presented the results of the first shared task on word sense induction for a Slavic language. For this shared task, three new large-scale datasets for word sense induction and disambiguation for the Russian language have been created and published. The shared task attracted 18 participating teams, which submitted overall 383 model runs on these three datasets. A substantial amount of the participant were able to outperform a competitive state-of-the-art baseline approach put in place by the organizers based on the AdaGram word sense embeddings method [Bartunov et al., 2016]. This shared task is the first systematic attempt to evaluate word sense induction and disambiguation systems for the Russian language. We hope that the produced resources and datasets will foster further research and help development of new generation of the sense representation methods.

Acknowledgements

This research was supported by Deutsche Forschungsgemeinschaft (DFG) under the projects JOIN-T and ACQuA and by the RFBR under the project no. 16-37-00354 мол_а. The work of Konstantin Lopukhin and Anastasiya Lopukhina was supported by a grant of the Russian Science Foundation, Project 16-18-02054. We are grateful to the authors of the Active Dictionary of Russian who kindly allowed us to use the dictionary in this shared task. We are grateful to Ted Pedersen for a helpful discussion about the evaluation settings. Finally, we thank our supporters and information sponsors, who helped to spread the word about the task: ABBYY, The Special Interest Group on Slavic Natural Language Processing (ACL SIGSLAV), Moscow Polytechnic University, Mathlingvo, and NLPub.

References

1. *Agirre E., Soroa A.* (2007), Semeval-2007 Task 02: Evaluating Word Sense Induction and Discrimination Systems, Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval '07), Prague, Czech Republic, pp. 7–12.
2. *Apresjan Ju. D.* (ed.) (2014), Active Dictionary of Russian. Vol. 1 (A–B), Vol. 2 (V–G) [Aktivnyj slovar' russkogo jazyka. Tom 1 (A-B), tom 2 (V–G)]. Jazyki slavjanskih kul'tur, Moscow, Russia.
3. *Apresjan V., Galaktionova I., Iomdin B.* (eds.). (2017), Active Dictionary of Russian. Vol. 3 (D–Z) [Aktivnyj slovar' russkogo jazyka. Tom 3 (D–Z)]. Nestor-Istoriya, Saint Petersburg, Russia.
4. *Arefyev N., Panchenko A., Lukanin A., Lesota O., Romanov P.* (2015), Evaluating three corpus-based semantic similarity systems for Russian, Computational Linguistics and Intellectual Technologies: papers from the Annual conference

"Dialogue" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Po Materialam Ezhegodnoy Mezhdunarodnoy Konferentsii «Dialog»], Vol. 2, Bekasovo, Russia, pp. 106–118.
5. *Arefyev N., Ermolaev P., Panchenko A.* (2018), How much does a word weigh? Weighting word2vec for word sense induction. Computational Linguistics and Intellectual Technologies: Papers from the Annual conference "Dialogue" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Po Materialam Ezhegodnoy Mezhdunarodnoy Konferentsii «Dialog»], Vol. 1, Moscow, Russia.
6. *Bartunov, S., Kondrashkin, D., Osokin, A., Vetrov, D. P.* (2016), Breaking Sticks and Ambiguities with Adaptive Skip-gram, Journal of Machine Learning Research, Vol. 51, pp. 130–138.
7. *Bojanowski P., Grave E., Joulin A., Mikolov T.* (2017), Enriching Word Vectors with Subword Information, Transactions of the Association for Computational Linguistics, Vol. 5, pp. 135–146.
8. *Hubert L., Arabie P.* (1985), Comparing Partitions, Journal of Classification, Vol. 2, № 1, pp. 193–218.
9. *Jurgens D., Klapaftis I.* (2013), SemEval-2013 Task 13: Word Sense Induction for Graded and Non-Graded Senses, Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, Georgia, USA, pp. 290–299.
10. *Kilgarriff A., Rychly P., Smrz P., Tugwell D.* (2004), The Sketch Engine, EuraLex 2004 Proceedings, pp. 105–115.
11. *Klein D. K., Murphy G. L.* (2001), The Representation of Polysemous Words, Journal of Memory and Language, Vol. 45, № 2, pp. 259–282.
12. *Klepousniotou E.* (2002), The Processing of Lexical Ambiguity: Homonymy and Polysemy in the Mental Lexicon, Brain and Language, Vol. 81, № 1–3, pp. 205–223.
13. *Kobritsov B. P., Lashevskaja O. N., Shemanaeva O. Yu.* (2005), Shallow Rules for Word-Sense Disambiguation in Text Corpora, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2005" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii «Dialog 2005»], Zvenigorod, Russia.
14. *Krippendorff K.* (2013), Content Analysis: An Introduction to Its Methodology (3rd Edition), SAGE, Thousand Oaks, CA, USA.
15. *Kutuzov A., Andreev I.* (2015), Texts in, meaning out: neural language models in semantic similarity task for Russian, Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Po Materialam Ezhegodnoy Mezhdunarodnoy Konferentsii «Dialog»], Vol. 2, Bekasovo, Russia, pp. 133–144.
16. *Kutuzov A.* (2018), Russian word sense induction by clustering averaged word embeddings: participation in RUSSE'2018 shared task. Computational Linguistics and Intellectual Technologies: Papers from the Annual conference "Dialogue" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Po Materialam Ezhegodnoy Mezhdunarodnoy Konferentsii «Dialog»], Vol. 1, Moscow, Russia.

17. *Kuznetsov S. A.* (ed.) (2014), Large Explanatory Dictionary of Russian [Bol'shoy Tolkoviy Slovar' Russkogo Yazika], available at: http://gramota.ru/slovari/info/bts/.
18. *Li J., Jurafsky D.* (2015), Do Multi-Sense Embeddings Improve Natural Language Understanding?, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015), Lisbon, Portugal, pp. 1722–1732.
19. *Lopukhin K., Lopukhina A.* (2016), Word Sense Disambiguation for Russian Verbs Using Semantic Vectors and Dictionary Entries, Computational Linguistics and Intellectual Technologies: Papers from the Annual conference "Dialogue" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Po Materialam Ezhegodnoy Mezhdunarodnoy Konferentsii «Dialog»], Moscow, Russia, pp. 393–405.
20. *Lopukhin K. A., Iomdin B. L., Lopukhina A. A.* (2017), Word Sense Induction for Russian: Deep Study and Comparison with Dictionaries, Computational Linguistics and Intellectual Technologies: Papers from the Annual conference "Dialogue" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Po Materialam Ezhegodnoy Mezhdunarodnoy Konferentsii «Dialog»], Vol. 1, Moscow, Russia, pp. 121–134.
21. *Lopukhina A., Lopukhin K., Nosyrev G.* (2018), Automated word sense frequency estimation for Russian nouns, Quantitative Approaches to the Russian Language, Routledge, Taylor & Francis Group, London, UK, pp. 79–94.
22. *Loukachevitch N., Chuiko D.* (2007), Thesaurus-based Word Sense Disambiguation [Avtomaticheskoe razreshenie leksicheskoy mnogoznachnosti na baze tezaurusnykh znaniy], Proceedings of the Contest "Internet Mathematics 2007" [Sbornik rabot uchastnikov konkursa «Internet-Matematika 2007»], Yekaterinburg, Russia, pp. 108–117.
23. *Loukachevitch, N., Chetviorkin I.* (2015), Determining the Most Frequent Senses Using Russian Linguistic Ontology RuThes, Proceedings of the workshop on Semantic resources and semantic annotation for Natural Language Processing and the Digital Humanities at NODALIDA 2015, Vilnius, Lithuania, pp. 21–27.
24. *Lyashevskaja O., Mitrofanova O.* (2009), Disambiguation of Taxonomy Markers in Context: Russian Nouns, Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009), Odense, Denmark, pp. 111–117.
25. *Lyons J.* (1977), Semantics, Vol. 2, Cambridge University Press, Cambridge, UK.
26. *Manandhar S., Klapaftis I., Dligach D., Pradhan S.* (2010), SemEval-2010 Task 14: Word Sense Induction & Disambiguation, Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval '10), Los Angeles, CA, USA, pp. 63–68.
27. *Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J.* (2013), Distributed Representations of Words and Phrases and their Compositionality, Advances in Neural Information Processing Systems 26, Harrahs and Harveys, NV, USA, pp. 3111–3119.
28. *Navigli R., Vannella D.* (2013), SemEval-2013 Task 11: Word Sense Induction and Disambiguation within an End-User Application, Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the

Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, GA, USA, pp. 193–201.
29. *Panchenko A., Ustalov D., Arefyev N., Paperno D. Konstantinova N., Loukachevitch N. and Biemann C.* (2016): Human and Machine Judgements about Russian Semantic Relatedness. In Proceedings of the 5th Conference on Analysis of Images, Social Networks, and Texts (AIST'2016). Communications in Computer and Information Science (CCIS). Springer-Verlag Berlin Heidelberg
30. *Pelevina M., Arefyev N., Biemann C., Panchenko A.* (2016): Making Sense of Word Embeddings. In Proceedings of the 1st Workshop on Representation Learning for NLP co-located with the ACL conference. Berlin, Germany. Association for Computational Linguistics
31. *Sharoff S.* (2006), Creating general-purpose corpora using automated search engine queries, WaCky! Working papers on the Web as Corpus, Gedit, Bologna, Italy, pp. 63–98
32. *Ustalov D., Panchenko A., Biemann C.* (2017), Watset: Automatic Induction of Synsets from a Graph of Synonyms, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, pp. 1579–1590.

ИЛЛОКУТИВНОЕ УПОТРЕБЛЕНИЕ СОЮЗОВ: ШКАЛА ИЛЛОКУТИВНОСТИ И ЕЕ ОТРАЖЕНИЕ В ГРАММАТИКЕ[1]

Пекелис О. Е. (opekelis@gmail.com)

Российский государственный гуманитарный университет, Москва, Россия

В статье рассматривается иллокутивное употребление союзов, при котором союз связывает пропозицию одной клаузы с иллокутивной модальностью другой. Обосновывается шкалярный подход к интерпретации этого явления: наряду с бесспорно иллокутивным и бесспорно неиллокутивным употреблением, существует класс конструкций с промежуточными свойствами. Формулируются критерии разграничения степеней иллокутивности. Демонстрируется, в частности, что императивные предложения, в отличие от вопросительных, не бывают бесспорно иллокутивными. Предъявляются свидетельства того, что предлагаемый подход находит подтверждение в грамматике: разные союзы совместимы с разными видами иллокутивного употребления; в составе бесспорно иллокутивных конструкций не употребляется коррелят *тогда*.

Ключевые слова: иллокутивное употребление, союз, иллокутивная модальность, императив, вопрос, подчинение, сочинение, коррелят

SPEECH ACT CONJUNCTION: THE SCALE OF SPEECH ACT USE AND ITS MANIFESTATION IN GRAMMAR

Pekelis O. E. (opekelis@gmail.com)

Russian State University for the Humanities, Moscow, Russia

This paper deals with the phenomenon of speech act conjunction in which the relation expressed by the conjunction holds on the level of speech act performance rather than on the level of states of affairs. It is argued that besides clearly speech act and clearly non-speech act uses, there is a class of constructions of an intermediate nature. The criteria are proposed

[1] Работа выполнена при поддержке гранта РФФИ № 16-06-00226.

that serve to distinguish between these three types of use. In particular, it is demonstrated that imperative sentences can only be of the “intermediate” type, while interrogative sentences can represent the clearly speech act use. The proposed distinction manifests itself in grammar. Namely, different conjunctions are compatible with different types of speech act use; the correlative item *togda* (‘then’) cannot be used within a clearly speech act construction.

Key words: speech act conjunction, illocutionary force, imperative, interrogative, subordination, coordination, correlative

1. Введение

Иллокутивным (англ. *speech act*) называют такое употребление союза, при котором союз выражает отношение между значением одной клаузы в составе сложного предложения и речевым актом, соответствующим другой клаузе. Говоря более формально, такой союз связывает пропозициональное содержание одной клаузы с иллокутивной модальностью другой ([Падучева 1985: 46], [Иорданская 1988], [Sweetser 1990], [Иорданская, Мельчук 2007: 430] и др.). Так, в (1) союз *чтобы* выражает целевую связь между пропозицией придаточного и модальностью утверждения, входящей в смысл главной клаузы (‘Чтобы быть до конца честным, *скажу*: дети в основном представлены девочками’)[2]:

(1) *<Дети задумчивы, понятливы, рисуют картинки на сложные библейские сюжеты.> Чтобы быть до конца честным, дети в основном представлены девочками.* [Игорь Мартынов. Дерибрюхово (1997) // «Столица», 1997.03.04]

Для сравнения, в (2) тот же союз *чтобы* соединяет пропозиции двух клауз, т. е. употреблен не иллокутивно (‘убрал для того, чтобы быть честным’)[3]:

(2) *<Курчев с радостью убрал ее в гардероб.> Чтобы быть совсем честным, убрал туда же и Марьянин клетчатый чемодан.* [Владимир Корнилов. Демобилизация (1969–1971)]

Хотя само понятие иллокутивного употребления союза общепринято, вопрос о его объеме и границах остается дискуссионным. Важная проблема, сформулированная в [Санников 2008: 66], состоит в том, что во многих случаях предложение с иллокутивно употребленным союзом оказывается близко по смыслу

[2] Примеры с указанием источника здесь и далее, если не сказано иное, заимствованы из Национального корпуса русского языка [НКРЯ].

[3] Указанному различию между (1) и (2) сопутствует ряд других различий. Так, в (1) субъектом целеполагания является говорящий, тогда как в (2) субъект целеполагания совпадает с субъектом главной клаузы. (Мы признательны рецензенту «Диалога», указавшему на необходимость этого уточнения.) Рамки статьи не позволяют остановиться на возможных сопутствующих признаках иллокутивного употребления подробнее.

предложению с неиллокутивным союзом. Так, предложения (3а) и (3б) близки по смыслу: и зависимые, и главные клаузы в их составе организованы одинаковыми предикатами и выражают похожие ситуации. При этом, однако, (3а) допускает иллокутивную интерпретацию (‘по причине того, что сахара нет, *я прошу тебя* варить на ксилите’), а (3б) ее не допускает (‘по причине того, что сахара нет, я #(*говорю, что*) буду варить на ксилите’). Тем самым, с одной стороны, (3а) удовлетворяет представлениям об иллокутивном употреблении, а с другой стороны, нет понятных оснований усматривать в (3а) иной тип употребления союза *раз*, чем в (3б).

(3) а. *Раз сахара нет, вари на ксилите.* [Татьяна Тронина. Русалка для интимных встреч (2004)]

б. *Раз сахара нет, буду варить на ксилите.*

В [Санников 2008: 56] критерием разграничения иллокутивных и неиллокутивных употреблений предложено считать наличие смысловой связи между частями сложного предложения: только в отсутствие такой связи можно говорить об иллокутивном употреблении союза. В соответствии с этим критерием, в примере (3а) союз *раз* употреблен не иллокутивно, поскольку налицо смысловая связь между пропозициями клауз.

Основания такого подхода понятны: в отсутствие смысловой связи между пропозициями неиллокутивное прочтение предложения, при котором союз как раз связывает пропозиции, оказывается невозможным. Тем не менее, по ряду причин этот подход не кажется исчерпывающим.

Во-первых, наличие или отсутствие смысловой связи не всегда легко диагностируется. Скажем, ответ на вопрос о наличии смысловой связи между пропозициями ‘быть честным’ и ‘дети представлены девочками’ в примере (1) представляется неочевидным.

Во-вторых, даже если согласиться с определением иллокутивного употребления на основании указанного трудноуловимого признака, такому определению не хватает грамматического «фундамента»: свидетельств того, что разграничение иллокутивно и неиллокутивно употребленных союзов по признаку наличия/отсутствия смысловой связи между пропозициями релевантно для русской грамматики.

Наконец, в третьих, указанным подходом охватываются не все проблематичные предложения. Так, в примере (4), с одной стороны, связь между пропозициями клауз как будто наличествует: под «светом в конце тоннеля» подразумевается положение дел с коррупцией.

(4) *Раз мы затронули тему коррупции, есть ли, на ваш взгляд, свет в конце тоннеля?* [Мы просто стараемся работать добросовестно в интересах наших клиентов (2004) // «Управление персоналом», 2004.11.15]

С другой стороны, этот пример невозможно интерпретировать как неиллокутивный. В самом деле, при неиллокутивном прочтении обе клаузы входят в сферу действия вопросительного оператора, а причинный союз *раз* соединяет две пропозиции, что дает семантически аномальный смысл: *#‘Я спрашиваю:*

есть ли свет в конце тоннеля по причине того, что мы заговорили о коррупции?'. Ср. с осмысленной иллокутивной интерпретацией: 'По причине того, что мы заговорили о коррупции, *я спрашиваю*: есть ли свет в конце тоннеля?'.

В настоящей работе предлагается иной подход к разграничению иллокутивных и неиллокутивных предложений[4]. Мы стремимся показать, что это разграничение не может быть описано как бинарное: на шкале иллокутивности имеется, наряду с полюсами — случаями бесспорно иллокутивного и бесспорно неиллокутивного употребления — класс промежуточных случаев. Ниже формулируются критерии разграничения бесспорного и промежуточного употреблений (**раздел 2**). Демонстрируется, в частности, что императивные и вопросительные конструкции ведут себя по-разному: первые не бывают бесспорно иллокутивными. Предлагается объяснение такого различия между императивом и вопросом (**раздел 3**). Наконец, предъявляются свидетельства того, что разграничение бесспорно неиллокутивных, бесспорно иллокутивных и промежуточных конструкций имеет под собой грамматические основания (**раздел 4**).

2. К определению шкалы иллокутивности

Для простоты далее рассматривается только такое иллокутивное употребление, при котором подчинительный союз соединяет пропозицию придаточного с иллокутивной модальностью главной клаузы (ср. пример (1)). Такая конфигурация, хотя и не является единственно возможной, для иллокутивных конструкций наиболее типична[5].

Сформулируем критерии, позволяющие разграничивать бесспорно иллокутивные, бесспорно неиллокутивные и промежуточные конструкции.

К **бесспорно иллокутивным** кажется оправданным относить такие конструкции, для которых а) иллокутивное и неиллокутивное прочтение существенно различаются по смыслу, при этом б) неиллокутивная интерпретация сомнительна.

Поясним, что под «существенным» смысловым различием мы понимаем прежде всего различие в терминах условий истинности. Так, в примере (4) осмысленной иллокутивной и аномальной неиллокутивной интерпретации

---

[4] Тем самым, мы пересматриваем позицию, изложенную ранее в [Пекелис 2013], где описание иллокутивного употребления союзов развивает подход, предложенный В. З. Санниковым.

[5] Вне рассмотрения, тем самым, оказываются две конфигурации: 1) соединение посредством союза иллокутивной модальности придаточного с пропозицией главной клаузы и 2) иллокутивное употребление сочинительного союза. Первая конфигурация, как правило, не может реализоваться, потому что у зависимой клаузы обычно не может быть собственной иллокутивной силы, ср. невозможность **Он испугался, когда что увидел?* при допустимости вопроса в составе сочиненной клаузы *Он вошел, и что он там увидел?* Иллокутивное употребление сочинительного союза встречается, ср. *извините, но вы ничего не понимаете* (≈'Извините, но *я скажу:* вы ничего не понимаете'), однако явно распространено менее, чем иллокутивное употребление подчинительных союзов. Коротко про случай сочинения см. в разделе 3.

соответствуют разные условия истинности: только для первой в условия истинности входит наличие причинной связи между вопросом *Есть ли свет в конце тоннеля?* и разговором о коррупции[6].

К **бесспорно неиллокутивным** относятся такие конструкции, для которых а) иллокутивное и неиллокутивное прочтение существенно различаются по смыслу, при этом б) иллокутивная интерпретация сомнительна.

Наконец, к **промежуточным** конструкциям мы относим такие, для которых а) иллокутивное и неиллокутивное прочтение существенно не различаются по смыслу, при этом б) оба прочтения осмысленны[7].

Проиллюстрируем каждое из определений примерами.

В (5) и (6) представлены бесспорно иллокутивные предложения, различающиеся типом иллокутивной модальности главной клаузы — утверждение в (5), вопрос в (6):

(5) *Если хочешь знать, крокодил умнее твоей собаки*. [М. С. Аромштам. Мохнатый ребенок (2010)]

(6) *<Не уверен, что мы говорим об одной и той же истории…> Но поскольку зашла речь, знаешь ли ты, ведьма, почему у драмы мужская маска, а у комедии — женская?* [Иржи Грошек. Реставрация обеда (2000)]

В самом деле, и (5), и (6) удовлетворяют данному выше определению бесспорно иллокутивных конструкций. В (5) иллокутивная интерпретация (‘При условии, что ты хочешь об этом знать, *я скажу тебе*: крокодил умнее твоей собаки’) и неиллокутивная (*#‘Я говорю тебе*: крокодил умнее твоей собаки при условии, что ты хочешь об этом знать’) существенно различаются по смыслу — имеют разные условия истинности, при этом вторая семантически аномальна[8]. То же имеет место и в (6) (ср. иллокутивную интерпретацию ‘По причине того, что

---

[6] Мы не углубляемся в трудный вопрос о том, что считать условиями истинности (или аналогом условий истинности) для вопросительного предложения, каковым при неиллокутивной интерпретации окажется пример (4) (см. на эту тему, например, [Hamblin 1958], [Partee 1991: 171]). Для наших целей достаточно, что совпадение в этих терминах иллокутивной и неиллокутивной интерпретаций (4) представляется заведомо невозможным, поскольку не сопоставимы условия истинности для входящих в их состав причинных пропозиций (‘есть свет в конце тоннеля по причине того, что заговорили о коррупции’ vs. ‘спрашиваю о том, есть ли свет в конце тоннеля, по причине того, что заговорили о коррупции’).

[7] Логически возможен четвертый вариант: обе интерпретации приемлемы и существенно различаются по смыслу. Ср. предложение: *Чтобы тебя успокоить, Вика приезжает завтра*. И иллокутивная интерпретация этого предложения ‘Чтобы тебя успокоить, *скажу*: Вика приезжает завтра’, и неиллокутивная ‘Вика приезжает завтра с целью тебя успокоить’ осмысленны. По-видимому, в примерах такого рода следует усматривать омонимию между бесспорно иллокутивным и бесспорно неиллокутивным прочтением, поскольку разные прочтения уместны в разных контекстах.

[8] Здесь и далее при описании неиллокутивного прочтения утвердительных конструкций, таких как (5), мы эксплицируем модальность утверждения, характеризующую все сложное предложение. Этим обеспечивается единообразный способ описания, во-первых, иллокутивных и неиллокутивных прочтений утвердительных конструкций, и во-вторых, конструкций с разными иллокутивными модальностями.

зашла речь, *я спрашиваю тебя:* знаешь ли ты, что…' с неиллокутивной и семантически аномальной #'*Я спрашиваю тебя*: знаешь ли ты, почему у драмы мужская маска, а у комедии — женская, по причине того, что об этом зашла речь?').

Третья основная разновидность иллокутивной модальности, императив, в бесспорно иллокутивных конструкциях за редкими исключениями не употребляется (см. подробнее раздел 3).

В (7), (8) и (9) представлены бесспорно неиллокутивные конструкции, главная клауза которых имеет иллокутивную модальность утверждения, императива и вопроса, соответственно:

(7) *Поскольку английский знаю недостаточно хорошо, приходится писать по-русски…* [коллективный. Форум: Похороните меня за плинтусом. Фильм (2009–2011)]

(8) *Повтори, пока не забыл: я запишу.* [Д. С. Мережковский. Смерть богов. Юлиан Отступник (1895)]

(9) *Если бы Ёжик с Медвежонком были вдвоём, зачем бы им понадобилось ещё три чашки?* [Сергей Козлов. Новогодняя сказка // «Мурзилка», 2003]

В (7) неиллокутивное прочтение '*Я говорю*: приходится писать по-русски по той причине, что английский знаю недостаточно хорошо' более осмысленно, чем иллокутивное '*Я говорю, что* приходится писать по-русски, по той причине, что английский знаю недостаточно хорошо'. Аналогично в (8) союз *пока* призван ограничить во времени ситуацию повтора, но не просьбу говорящего повторить. Так же и в (9) иллокутивная интерпретация 'Если бы Ежик с Медвежонком были вдвоем, *я спрашиваю/спросил бы*: зачем бы им понадобилось ещё три чашки?' несостоятельна против неиллокутивного прочтения '*Я спрашиваю*, зачем бы им понадобились три кружки при условии, что они были бы вдвоем?'.

Наконец, примеры (10)–(12) иллюстрируют промежуточную конструкцию для сообщения, императива и вопроса, соответственно:

(10) *Поэтому, раз уж вы временно занимаете мое место, я хочу вас использовать.* [Татьяна Устинова. Персональный ангел (2002)]

(11) *Если вы христианин, попробуйте ответить на мой вопрос: оружие, оно от Бога?* [коллективный. Форум: Горный двухподвесочный (2010)]

(12) *Раз Надежда Васильевна любит Наташку, почему бы Наташке не полюбить Надежду Васильевну?* [Владимир Железников. Жизнь и приключения чудака (1974)]

В соответствии с данным выше определением, в (10), (11) и (12) равно приемлемы иллокутивное и неиллокутивное прочтение, причем видимой разницы в значении (в условиях истинности) между двумя прочтениями нет. Ср. иллокутивное прочтение 'По причине того, что вы временно занимаете мое место, *я говорю,* что хочу вас использовать' и неиллокутивное '*Я говорю*, что по причине того, что вы временно занимаете мое место, я хочу вас использовать' для (10); иллокутивное 'При условии, что вы христианин, *я прошу вас* попробовать

ответить на мой вопрос' и неиллокутивное '*Я прошу,* чтобы при условии, что вы христианин, вы попробовали ответить на мой вопрос' для (11); иллокутивное 'По причине того, что Надежда Васильевна любит Наташку, *я спрашиваю*: почему бы Наташке не полюбить Надежду Васильевну?' и неиллокутивное '*Я спрашиваю*: почему бы Наташке не полюбить Надежду Васильевну по причине того, что Надежда Васильевна любит Наташку?' для (12)[9].

3. Об особой позиции императива на шкале иллокутивности

Сложноподчиненное предложение с матричным глаголом в императиве, по-видимому, не может получить бесспорно иллокутивной интерпретации. Другими словами, любое такое предложение, если оно допускает иллокутивную интерпретацию, допускает и близкую по смыслу неиллокутивную. Ср. пример (11), приведенный выше, а также следующую пару примеров, различающихся тем, что вопросительное предложение (13) является бесспорно иллокутивным, тогда как императивное предложение (14) осмысленно в обеих интерпретациях:

(13) *Раз уж мы заговорили о читателях, что ты о них знаешь?* [Людмила Палисад. Евгений Беркович: «Пока издание интересно читателям и мне, силы находятся» (2003) // «Вестник США», 2003.12.10] — иллокутивная интерпретация: 'По причине того, что мы заговорили о читателях, *я спрашиваю тебя*, что ты о них знаешь?'; неиллокутивная интерпретация: #'*Я спрашиваю тебя*, что ты знаешь о читателях по причине того, что мы о них заговорили?'.

(14) *Раз, Коля, там будет бомонд, надевай новые джинсы.* [Григорий Горин. Сауна (1974–1984)] — иллокутивная интерпретация: 'По причине того, что там будет бомонд, *я прошу тебя* надеть новые джинсы'; неиллокутивная интерпретация: '*Я прошу тебя,* чтобы по причине того, что там будет бомонд, ты надел новые джинсы'.

Высказанное предположение подтверждается некоторыми грамматическими различиями между конструкцией с императивом и вопросительной конструкцией (см. **раздел 4**).

Причина указанного отличия императива от вопроса видится в том, что взаимодействие импликативной семантики союза с императивной иллокутивной

[9] По мнению рецензента «Диалога», излагаемый подход к разграничению иллокутивного и неиллокутивного употреблений трудно формализовать, а значит, трудно верифицировать. С этим можно согласиться лишь отчасти. То, что языковые единицы могут поразному себя вести в зависимости от того, с каким уровнем синтаксической структуры они взаимодействуют — с уровнем, соответствующим речевому акту, пропозиции или глагольной группе — описано в формальном синтаксисе; см., например, в [Krifka 2013] объяснение в этих терминах свойств слов *yes* и *no*. Вместе с тем, поскольку в рамках настоящей работы задача формализации не ставится, то и говорить о ее решаемости преждевременно. Сказанное не означает, однако, что развиваемый подход мы оставляем без аргументации: доводом в его пользу видится отражение предполагаемых семантических противопоставлений в грамматике, о котором см. **раздел 4**.

модальностью с точки зрения условий истинности обычно мало отличается от аналогичного взаимодействия союза с пропозицией, ассоциируемой с императивной модальностью. Так, в примере (14) причинный смысл, выражаемый союзом, может взаимодействовать с иллокутивной модальностью императива (‘прошу <надеть> по причине того, что…’), а может — с пропозициональным содержанием, ассоциируемым с императивной формой (‘чтобы надел по причине того, что…’). Если первый смысл выражает истинное положение дел, то и второй обычно не является ложным. Шире, если говорящий X, ссылаясь на причину Z, просит слушающего Y выполнить действие P, в общем случае верно, что X просит, чтобы Y выполнил P, руководствуясь причиной Z.

Напротив, взаимодействие импликативной семантики союза с вопросительной иллокутивной модальностью вовсе не гарантирует, что осмысленно и совпадает по условиям истинности аналогичное взаимодействие между союзом и пропозицией, ассоциированной с вопросом (ср. в (13) ‘спрашиваю по причине того, что…’ и ‘знаешь по причине того, что…’). Таким образом, особое положение императива, в конечном счете, обусловлено особенностью императивной семантики: способностью императива «транслировать» содержание своего логико-семантического взаимодействия с внешним контекстом соответствующей пропозиции.

О том, что отмеченный эффект определяется прежде всего семантикой императива, а не собственно грамматической формой, свидетельствуют примеры типа (15). В (15а) каузативная семантика выражена индикативной формой, при этом соотнесение причинного смысла с вершинным предикатом (‘по причине того, что ненавижу летать, прошу, чтобы вы выпили’) и с вложенным (‘прошу, чтобы по причине того, что я ненавижу летать, вы выпили’) дает интерпретации, близкие с точки зрения условий истинности. Близость двух интерпретаций иллюстрирует и пример (15б), полученный трансформацией (15а) в конструкцию с императивом[10]:

(15) а. *Я прошу вас выпить со мной, потому что я ненавижу летать.*
[Мария Голованивская. Противоречие по сути (2000)]

---

[10] Исключение составляет индикативная форма типа *прошу*, употребленная не перформативно. Ср. (i), где форма *прошу* (по крайней мере, в одной из интерпретаций) не равносильна осуществлению соответствующего действия — просьбы:

(i) *Я тебя прошу пойти на встречу, потому что искренне боюсь за тебя*. [Алексей Иванов. Комьюнити (2012)]

В этом случае два обсуждаемых прочтения (‘из-за того, что я боюсь за тебя, я прошу тебя пойти на встречу’ и ‘я прошу, чтобы ты пошла на встречу из-за того, что я боюсь за тебя’) имеют, по-видимому, разные условия истинности (вопрос о том, каковы причины этого отличия неперформативных высказываний с *прошу* от перформативных, мы оставляем за рамками работы).

Данное исключение, однако, на рассматриваемые в работе конструкции с императивом не распространяется. В самом деле, высказывания, организуемые императивной формой, всегда перформативны [Добрушина 2014]. Закономерным образом, трансформация (i) в конструкцию с императивом затруднена:

(ii) ?*Пойди на встречу, потому что я искренне боюсь за тебя*.

б. *Выпейте со мной, потому что я ненавижу летать.*

Следует отметить, что недоступность бесспорно иллокутивной интерпретации касается только рассматриваемых здесь сложноподчиненных предложений с главным глаголом в императиве. Сложносочиненные предложения с формой императива в одной из клауз бывают бесспорно иллокутивными. Ср. следующий пример, заимствованный в работе [Падучева 1985: 46][11]:

(16) *Хлеба тоже нет, так что зайди в булочную.*

Одно из отличий сочинения от подчинения принято усматривать в том, что в сочинительной полипредикативной конструкции каждой из клауз соответствует своя иллокутивная сила (см. [Cristofaro 2003], [Verstraete 2005] и др.). Поэтому неиллокутивная интерпретация, при которой в сфере действия императива оказывается все сложное предложение, в этом случае невозможна (ср. *#'Я прошу тебя*, хлеба нет, поэтому зайти в булочную'). В качестве единственно допустимой остается иллокутивная интерпретация: 'Хлеба тоже нет, поэтому *я прошу тебя* зайти в булочную'.

4. Шкала иллокутивности и ее грамматические проявления

В настоящем разделе рассмотрены два явления, позволяющие говорить о том, что предложенное разграничение степеней иллокутивности отражается в грамматике: зависимость между степенью иллокутивности и выбором союза (**раздел 4.1**) и различная сочетаемость иллокутивных конструкций с коррелятом *тогда* (**раздел 4.2**).

4.1. Разные союзы на шкале иллокутивности

Союзы обнаруживают различия с точки зрения способности употребляться в бесспорно иллокутивных, бесспорно неиллокутивных и промежуточных конструкциях. Так, союзы *чтобы, раз* и *если* способны организовывать бесспорно иллокутивную конструкцию (ср. примеры (1), (4) и (5)). Союз *потому что*, между тем, к бесспорно иллокутивному употреблению не способен, но может употребляться в промежуточной конструкции[12]. Так, модификация бесспорно иллокутивного примера (4), с заменой союза *раз* на *потому что*, дает неприемлемый результат (17), тогда как замена *раз* на *потому что* в примере

[11] См. в [Пекелис 2015] обоснование того, что фигурирующий в (16) союз *так что* является сочинительным. Обратим внимание, кроме того, что форма императива обычно невозможна в подчинительном контексте [Добрушина 2014], ср. *так что зайди в булочную* vs. **потому что зайди в булочную.*

[12] Неспособностью к бесспорно иллокутивному употреблению *потому что* отличается от английского *because*, ср. пример из [Verstraete 1999]: *John is here, because I don't want you to run into him unprepared*. — букв. 'Джон здесь, потому что я не хочу, чтобы ты столкнулся с ним неожиданно'.

(14), содержащем императив и представляющем собой, тем самым, промежуточную конструкцию, допустима, ср. (18):

(17) **Есть ли, на ваш взгляд, свет в конце тоннеля, потому что мы затронули тему коррупции?*

(18) *Надевай новые джинсы, потому что там будет бомонд.*

Ср. также корпусный пример (20) с *потому что* и матричным глаголом в императиве.

Наконец, союз *оттого что* не допускает не только бесспорно иллокутивного, но и промежуточного употреблений. Ср. невозможность замены *потому что* на *оттого что* в (19) (=(18)) и (20):

(19) *Надевай новые джинсы, потому что <??оттого что> там будет бомонд.*

(20) *Старайся быть счастливой, потому что <??оттого что> жизнь одна и проходит быстро.* [Андрей Волос. Недвижимость (2000) // «Новый Мир», 2001]

Отметим, что в НКРЯ подобные примеры с *оттого что* и императивом отсутствуют.

4.2. Коррелят тогда как индикатор степени иллокутивности

Коррелятивный показатель *тогда* не употребляется в бесспорно иллокутивных конструкциях. Так, *тогда* неприемлем в составе бесспорно иллокутивных примеров (21) и (22) (модификация примеров (4) и (5), соответственно), но допустим в составе бесспорно неиллокутивного примера (23) (модифицированный (9)) и промежуточного (24):

(21) *??Раз мы затронули тему коррупции, тогда есть ли, на ваш взгляд, свет в конце тоннеля?*

(22) **Если хочешь знать, тогда крокодил умнее твоей собаки.*

(23) *ОКЕсли бы Ёжик с Медвежонком были вдвоём, тогда зачем бы им понадобилось ещё три чашки?*

(24) *Раз слышал, тогда иди и выполняй.* [Евгений Сухов. Делу конец — сроку начало (2007)]

Причина того, что *тогда* несовместим с бесспорной иллокутивностью, видится в том, что *тогда* служит средством контрастивного выделения придаточного [Podlesskaya 1997], тогда как в иллокутивной конструкции, наоборот, придаточное характеризуется низкой коммуникативной значимостью, выступая всего лишь средством обоснования речевого акта, выраженного в главной клаузе [Verstraete 1999].

5. Заключение

Подведем итог.

- Шкалу иллокутивного употребления союзов составляют два полюса — бесспорно иллокутивные и бесспорно неиллокутивные конструкции — и класс промежуточных случаев.
- Такое разграничение уместно основывать на двух параметрах: осмысленность иллокутивной и неиллокутивной интерпретации и наличие смыслового различия между интерпретациями.
- Конструкции с главным глаголом в императиве, в отличие от утвердительных и вопросительных конструкций, из-за особой семантики императива не бывают бесспорно иллокутивными.
- Разграничение трех степеней иллокутивности находит отражение в грамматике: разные союзы совместимы с разными степенями иллокутивности; коррелят *тогда* не допустим в составе бесспорно иллокутивных конструкций.

Литература

1. *Добрушина Н. Р.* (2014). Императив. Материалы для проекта корпусного описания русской грамматики (http://rusgram.ru). На правах рукописи. М.
2. *Иорданская Л. Н.* (1988). Семантика русского союза *раз* (в сравнении с некоторыми другими русскими союзами). Russian Linguistics, 12(3). С. 239–267.
3. *Иорданская Л. Н., Мельчук И. А.* (2007). Смысл и сочетаемость в словаре. М.: Языки славянских культур.
4. *Падучева Е. В.* (1985). Высказывание и его соотнесенность с действительностью. М.: Наука.
5. *Падучева Е. В.* (2016). Модальность. Материалы для проекта корпусного описания русской грамматики (http://rusgram.ru). На правах рукописи. М.
6. *Пекелис О. Е.* (2013). Иллокутивное употребление союзов. Материалы для проекта корпусного описания русской грамматики (http://rusgram.ru). На правах рукописи. М.
7. *Пекелис О. Е.* (2015). Сочинение и подчинение. Материалы для проекта корпусного описания русской грамматики (http://rusgram.ru). На правах рукописи. М.
8. *Санников В. З.* (2008). Русский синтаксис в семантико-прагматическом пространстве. М.: Языки славянских культур.
9. *Cristofaro S.* (2003). Subordination. Oxford: Oxford University Press.
10. *Hamblin, Ch. L.* (1958), Questions, The Australasian Journal of Philosophy, Vol. 36, pp. 159–168.
11. *Krifka, M.* (2013), Response particles as propositional anaphors, SALT: Proceedings of Semantics and Linguistic Theory, Snider, Todd (ed.), Vol. 23, pp. 1–18.
12. *Partee, B. H.* (1991), Topic, focus and quantification, SALT I: Proceedings of the First Annual Conference on Semantics and Linguistic Theory, eds. Moore S.,

Wyner A. Z., Ithaca, N. Y.: CLC Publications, Department of Linguistics, Cornell University, pp. 159–187.
13. *Podlesskaya V. I.* (1997). Syntax and semantics of resumption: some evidence from Russian conditional conjuncts. Russian Linguistics, Vol. 21 (2). Pp. 125–155.
14. *Sweetser E.* (1990). From Etymology to Pragmatics. Metaphorical and Cultural Aspects of Semantic Structure. Cambridge: Cambridge University Press.
15. *Verstraete J.-C.* (1999). The distinction between epistemic and speech act conjunction. Belgian Essays on Language and Literature. Pp. 119–130.
16. *Verstraete J.-C.* (2005). Two types of coordination in clause combining. Lingua, Vol. 115 (4). Pp. 611–626.

References

1. *Cristofaro S.* (2003), Subordination, Oxford University Press, Oxford.
2. *Dobrushina N. R.* Imperative [Imperativ], Towards a corpus description of Russian grammar [Materialy dlja proekta korpusnogo opisanija russkoj grammatiki], available at: (http://rusgram.ru), Manuscript, Moscow.
3. *Hamblin, Ch. L.* (1958), Questions, The Australasian Journal of Philosophy, Vol. 36, pp. 159–168.
4. *Iordanskaja L. N.* (1998), Semantics of the Russian conjunction *raz* (in comparison with other Russian conjunctions) [Semantika russkogo sojuza *raz* (v sravnenii s nekotorymi drugimi russkimi sojuzami)], Russian Linguistics, Vol. 12 (3), pp. 239–267.
5. *Iordanskaja L. N., Melchuk I. A.* (2008), Meaning and combinability in the dictionary [Smysl i sochetaemost' v slovare], Yazyki slavjanskih kul'tur, Moscow.
6. *Krifka, M.* (2013), Response particles as propositional anaphors, SALT: Proceedings of Semantics and Linguistic Theory, Snider, Todd (ed.), Vol. 23, pp. 1–18.
7. *Paducheva E. V.* (1985), Utterance and its relationship to reality [Vyskazyvanie i ego sootnesennost' s deistvitel'nost'ju], Nauka, Moscow.
8. *Paducheva E. V.* (2016), Modality [Modal'nost'], Towards a corpus description of Russian grammar [Materialy dlja proekta korpusnogo opisanija russkoj grammatiki], available at: (http://rusgram.ru), Manuscript, Moscow.
9. *Partee, B. H.* (1991), Topic, focus and quantification, SALT I: Proceedings of the First Annual Conference on Semantics and Linguistic Theory, eds. Moore S., Wyner A. Z., Ithaca, N. Y.: CLC Publications, Department of Linguistics, Cornell University, pp. 159–187.
10. *Pekelis O. E.* (2013), Speech act use of conjunctions [Illokutivnoe upotreblenie sojuzov], Towards a corpus description of Russian grammar [Materialy dlja proekta korpusnogo opisanija russkoj grammatiki], available at: (http://rusgram.ru), Manuscript, Moscow.
11. *Pekelis O. E.* (2015), Coordination and subordination [Sochinenije i podchineije], Towards a corpus description of Russian grammar [Materialy dlja proekta korpusnogo opisanija russkoj grammatiki], available at: (http://rusgram.ru), Manuscript, Moscow.

12. *Podlesskaya V. I.* (1997), Syntax and semantics of resumption: some evidence from Russian conditional conjuncts, Russian Linguistics, Vol. 21 (2), pp. 125–155.
13. *Sannikov V. Z.* (2008), Russian syntax in the semantic-pragmatic space [Russkij sintaksis v semantiko-pragmaticheskom prostranstve], Jazyki Slavjanskih Kul'tur, Moscow.
14. *Sweetser E.* (1990), From Etymology to Pragmatics. Metaphorical and Cultural Aspects of Semantic Structure, Cambridge University Press, Cambridge.
15. *Verstraete J.-C.* (1999), The distinction between epistemic and speech act conjunction, Belgian Essays on Language and Literature, pp. 119–130.
16. *Verstraete J.-C.* (2005), Two types of coordination in clause combining, Lingua, Vol. 115 (4), pp. 611–626.

SEMI-AUTOMATIC INTEGRATION OF A NEW LANGUAGE INTO A MULTILINGUAL NLP MODEL: THE CASE OF JAPANESE

Petrova M. A. (maria_pet@abbyy.com),
Druzhkina A. A. (anna_r@abbyy.com),
Garashchuk R. V. (ruslan_g@abbyy.com),
Yudina M. V. (maria_yu@abbyy.com)
ABBYY, Moscow, Russia

The current paper deals with the integration of the Japanese language in a multilingual NLP model, namely, the Compreno model. The formalism includes morphological, syntactic and semantic patterns, covering all possible semantic and syntactic dependencies a word can attach. The architecture of the model allows us to acquire nearly all semantic links of a word through its proper positioning in a thesaurus-like semantic hierarchy, where words are linked through semantic dependencies. The inheritance principle of the hierarchy simplifies the syntactic description of a newly added language as well. Unlike the traditional approach to Japanese parsing based on chunks, or bunsetsus, we suggest a Japanese parser based on constituents. Special attention is given to the tools that allow us to automatize language description process and significantly speed up the description. The work on the Japanese model is still in progress, therefore, we show the current results we have achieved, and point out problems that remain to be solved.

Keywords: Japanese parsing, multi-lingual parsing, semantic and syntactic analysis, formal language models, information extraction

ПОЛУАВТОМАТИЧЕСКАЯ ИНТЕГРАЦИЯ НОВОГО ЯЗЫКА В МНОГОЯЗЫЧНУЮ NLP-МОДЕЛЬ (НА ПРИМЕРЕ ЯПОНСКОГО ЯЗЫКА)

Петрова М. А. (maria_pet@abbyy.com),
Дружкина А. А. (anna_r@abbyy.com),
Гаращук Р. В. (ruslan_g@abbyy.com),
Юдина М. В. (maria_yu@abbyy.com)
ABBYY, Москва, Россия

1. Introduction and related work

The given paper is devoted to the integration of the Japanese language in the ABBYY Compreno model—NLP model based on morphological, syntactic and semantic text analysis. The model serves as the basis for a dependency parser and helps to apply text mining algorithms to different NLP tasks. Currently, it functions for English, Russian, German and, partly, for French, Spanish and Chinese.

Here we focus on linguistic problems bound with the process of automating language description, and show our experience of introducing new tools, which make the description semi-automatic and, therefore, more effective.

Japanese NLP tools are in high demand now, and there are quite a few works devoted to Japanese parsing ([Uchimoto et al. 2000]; [Kudo and Matsumoto 2002]; [Kurohashi and Nagao 1994]; [Kawahara and Kurohashi 2006]; [Kawahara et al. 2017]; [Tanaka, Nagata 2015], and others).

Traditionally, Japanese parsers differ significantly from the ones for European languages, as they are mostly based on syntactic dependencies modeled in terms of chunks called bunsetsus instead of constituents (the example of bunsetsus-based parsers are CaboCha [Kudo, Matsumoto 2002] and KNP [Kawahara, Kurohashi 2006]).

Taking the specificity of the Japanese language into account (especially, the problem of text division into words), this approach has some benefits. However, the bunsetsus-based models have significant disadvantages bound with the difficulties of setting correspondences between the bunsetsus and constituents. As [Tanaka, Nagata 2015: 237] points out, it "complicates the task of extracting semantic units from bunsetsus-based representations" and makes the analysis of non-tree links such as coordination problematic.

There are different studies aimed at improving parsing quality.

Recently, word-based dependency schemes have been suggested for Japanese, particularly within the UD project ([Nivre et al. 2016]; [Kanayama et al. 2015]; [Tanaka et al. 2016]). Besides, there are studies showing that adding lexical knowledge can significantly improve dependency analysis [Kawahara et al. 2017]. The use of case frames has also been reported to enhance parsing quality [Kawahara, Kurohashi 2006]; [Kawahara et al. 2017].

We suggest a different approach, in which the Japanese parser is based on a linguistic model, which includes not only lexical knowledge, but a full language description, covering all possible semantic and syntactic dependencies a word can attach. This is possible due to the integration of the Japanese vocabulary into the universal semantic model, which is much broader than the case frames application. Unlike bunsetsus-oriented parsers, our Japanese parser is based on constituents, which provides faster and easier integration of Japanese in a multilingual system.

The structure of the paper is as follows. First, we give a short description of the Compreno linguistic model in general, as far as it is necessary for further understanding (for more details see [Anisimovich et al. 2012; Manicheva et al. 2012; Petrova 2014]). Then we focus on integrating Japanese in the formalism and characterize the semantic and syntactic patterns of our Japanese description, drawing particular attention to the automation methods. Following this, we illustrate the work of the parser based on the given model—both for English and Japanese, and give a short

description of the corpora annotation used in the project. Finally, we offer the conclusion, where the results are summarized and further perspectives are given.

2. The Compreno linguistic model

The Compreno model is based on a multilingual lexical database organized in the form of a thesaurus-like hierarchy (Semantic Hierarchy, hereafter as SH)—a hyper-hyponymy relation tree built on a universal language, an interlingua. The branches of the tree are the so-called universal semantic classes (SCs)—universal labels, or boxes, which are filled with the contents in different natural languages. For instance, the tree includes the path such as PHYSICAL_OBJECT > BEING > ANIMAL > MYTHOLOGICAL_ANIMAL > DRAGON, and the DRAGON class is filled with the English 'dragon', German 'Drache', Japanese '竜', and so on.

The semantic links between words are provided through the *Deep Slots* (DSs)—semantic roles, under which we understand any semantic dependency a word can attach, like agent in '*[the cat] ran away*', or evaluation characteristic in *'a [nice] house'*.

The basic SH principle is the inheritance principle: all the DSs and other semantic features are introduced as high as possible in the hierarchy, and the lower branches inherit them. Such a strategy minimizes the amount of work necessary for the description of each word's semantic links: that is, when a new word is positioned in the hierarchy, it inherits nearly all possible semantic links it can have.

The SCs and the DSs are universal and do not depend on any definite language. It means that when we add a new language in the model, the semantic part of its description comes to efficient word positioning in the hierarchy, which provides the word with all the necessary DSs at once.

Each DS has a number of syntactic realizations—so called *Surface*, or *Syntactic*, *Slots* (SSs). Unlike the DSs, SSs are not universal. For each SS, we specify its grammar value—define the parts of speech that can fill it, indicate case, prepositions and other grammatical information, set its order in a sentence, and punctuation. For example, the Agent DS corresponds to the $Subject SS in '*[the boy] reads a book*' and to the $Object_Indirect_By slot—in the passive transformation '*the book is read [by the boy]*'.

The syntactic pattern of a newly added language demands more work, as syntax is special for each language. Although the inheritance principle helps here as well, we still face a great deal of work trying to determine, first, which surface realizations each DS can have, and, second, adding this information to the model.

3. Adding the Japanese language in the model

3.1. The semantic description: word positioning in the hierarchy

As shown above, proper word positioning in the SH is the key point of the semantic pattern, as it provides each word with an entire semantic model.

Previous work with other languages has proven that manual descriptions based on dictionaries are ineffective and take too much time. To overcome this, we developed a semi-automated (or semi-supervised) approach to adding new vocabulary, which was first used for the description of the German language (for details, see [Goncharova et al. 2015]). Using the approach, we created an auxiliary dictionary-like tool: on the one hand, it accumulated all relevant information from dictionaries and corpora—meanings of words, examples, and grammatical features; on the other hand, it automatically suggested a SC for each meaning of the word, and a linguist had only to approve or reject it.

In the current formalism, the number of the word meanings corresponds to the number of the SCs where a unique lexeme is represented. That is, each pair *lexeme—SC* represents one word meaning. We can get a number of such pairs for each language of the model. Moreover, we can get a frequency of each pair parsing parallel corpora with the Compreno parser, and, therefore, we obtain a variety of statistically ranged meanings.

When the work on the Japanese morphological system was completed, we aligned the Japanese-English parallel corpora and dictionaries with our alignment parser, found word pairs, where a certain Japanese word form corresponds to a certain English word form, and counted the frequency for each pair. As we have already had the mapping of the English word forms into SCs, we could obtain some hypotheses, or suggestions, for positioning Japanese words as well.

Therefore, we got a number of suggestions for each Japanese lexeme on where to place it in the SH and ranged them according to their frequency.

When we started using the tool for German, the percent of the correct suggestions in top 5 hypotheses was about 0.6 at first. By the time we started the Japanese description, the tool was significantly improved, and the algorithm switched from the heuristic-based approach to machine learning.

Namely, we evaluated the correctness of the German suggestions after the German vocabulary had been checked by linguists, and taught the system on it, as the classifier estimates the good and the bad features of the hypotheses (the features include word's/suggestion's part of speech, source of the suggestion (dictionary or parallel texts), distances between meanings in the source and target languages in the SH, depth of the suggestion in the SH, and other). As a machine learning method, gradient boosting over decision trees has been chosen.

Currently the algorithm gives us 0.72 precision within the top 5 results. It increased the speed of the Japanese description about 5 times in comparison with the German one, when the tool was used for the first time (we do not compare it with the speed of English and Russian, as their description was done together with elaborating the SH, DSs, SSs and other universal features of the system, which also took time).

Another option of the word positioning tool is to analyze, which additional semantic and grammatical features a word can have. It suggests not only the proper place for a word, but also other features, such as *semantemes* (universal features, which distinguish antonyms like *bad* and *good*, for instance) and *grammemes* (language-specific features, which describe the syntactic behaviour of a word (mark transitivity or government, for example).

Automatic suggestions like these come from several sources. First, some grammemes are shared within languages, like 'CharacteristicParametric' for parametric nouns; therefore, we assume that if English or Russian descendants of some SC have this grammeme, it is most likely that Japanese descendants can need it, too. Second, some suggestions are calculated from the models of the Japanese lexis already introduced in the hierarchy: if a Japanese word has some semanteme or grammeme, it is likely that its newly added neighbors will need them, too. Third, there are grammemes that are always relevant for some word groups,—for instance, all verbs must have a transitivity marker. Therefore, transitivity grammemes are always suggested when dealing with a verb.

A linguist now only has to test whether the positioning hypothesis is right, and if yes, to choose additional features from the list of the already generated suggestions. If the suggestion is incorrect, which is usually easy to find out from the information provided with the vocabulary tool (definition, different examples from the web, and so on), the correct SC can be chosen manually.

Currently we have more than 35,000 Japanese lexical units in the SH. For comparison, the total number of the universal SCs is about 190,000, and the number of English and Russian lexical units is nearly 270,000 and 247,000, correspondingly.

3.2. The syntactic description

Unlike semantics, syntax is special for every language. Therefore, when we add a new language in the model, we have to make a full description of its syntax. To make the work faster, we use the tools described below and turn to the inheritance principle again.

However, different dependencies demand different strategies. There are DSs that have the same syntactic realizations with every core they can be attached to, and the expression of some DSs depends on the cores they combine with.

That is, the syntactic realization of the adjuncts such as Purpose, Cause, or Condition does not depend on the verb they are bound with. For instance, every verb that can attach the reason slot can have ため-reason adjunct, as in example (1):

(1) 列車	は	[雪	の	ために]	遅れた	
ressya-train	wa-Nom	yuki-snow	no-Gen	tameni-Caus	okureta-be delayed-past	

The train was delayed [because of snow].

This means that we can indicate just once that the cause DS corresponds to the cause adjunct with the necessary grammatical properties. Therefore, the main task for the surface description of such DSs is to find all possible syntactic realizations for them.

To achieve it, we take parallel English-Japanese texts and analyze their English part with our parser in order to get all possible constituents corresponding to the necessary DS. In this way, we get parent-child pairs for each DS we need. After this, we find all possible Japanese correspondences for the lexemes that fill the DS. As Japanese is a left branching language, we check additionally that the supposed child node precedes the supposed parent node, and find the postposition closely following

the child. Therefore, we get a table-like catalogue of possible grammar realizations of each DS.

All grammar realizations are grouped into separate files according to realization markers. The files are ranged by the frequency of each realization: the larger the file, the more examples were found for this particular marker in the current search.

Each file contains a table, which provides: a) a source language instance with the parent node in red and the child node in blue (based on the default syntactic analysis by Compreno); b) a corresponding target language instance with the same colour code plus the marker in green; c) the vocabulary form for both nodes and the marker in the target language.

For instance, see Table 1—a small fragment of the file for the Cause DS expressed through the から postposition:

Table 1. The Cause DS expressed through the から postposition

source language instance	target language instance	the vocabulary form of parent and child nodes and the marker in the target language
I don't say that just because of your circumstances.	あなたの境遇から 言った訳ではない	言う から 境遇
He was called 'Eiki no oyakata' (Master in Eiki) because of his address.	住所から「永木の親方」と呼ばれた。	呼ぶ から 住所
For some reason, he grew up in a fatherless family.	家庭的な事情から、母子家庭で育つ。	育つ から 事情
Inventions are born, so to speak, of necessity.	発明はいわば必要から生まれるのだ。	生む から 必要
It is also called akoyamochi (lit 'oyster mochi') because of its shape.	その形からあこや餅とも呼ばれることもある。	呼ぶ から 形
Because of the importance of their role, they were allowed to adopt surnames and wear pairs of swords.	その役目の重要性から苗字帯刀を許されていた。	許す から 重要性

Nevertheless, the expression of some DSs depends on the core predicate. Mainly, this concerns the actant DSs, such as agent, object, experiencer, or alike. For example, in sentences (2)

(2) *He touched* [the water] *with his foot.*
I gave [a present] *to my friend.*

the bracketed constituent corresponds to the [Object] DS. The Japanese verbs 触れる [fureru] 'to touch' and 上げる [ageru] 'to give' have different government: 触れる has *ni*-Object (Dative Object) and 上げる demands *wo*-Object (Accusative Object). This means we must not only indicate that the [Object] DS can be expressed through *wo*-groups and *ni*-groups, but also indicate that different cores demand different syntactic realizations of the object-slot.

We describe this information in a semi-automated manner. First, we assign all possible realizations for the DSs like [Object]. Then we add grammemes for each type of the object-government, and provide the verbs with the necessary grammemes, namely, 触れる acquires the <NiObject>-grammeme and 上げる—the

<AccusativeObject>. All possible surface realizations of the [Object] DS are introduced high in the hierarchy, but the core of each correspondence is marked with the necessary grammeme: the *ni*-realization demands that the core verb should have the <NiObject>-grammeme, and the *wo*-realization demands the accusative grammeme.

When a verb is placed in the SH, relevant grammemes are suggested within the procedure described above.

This means that the syntactic description of most of the DSs, such as adjuncts and characteristics, is universal, so to say, as their syntactic realizations are introduced only once in the SH. The syntactic description of the DSs, which have lexicalized realization, is done in a semi-automated manner.

The model includes about 330 DSs. Currently, more than 70% of them are provided with Japanese surface realizations. However, the fullness of this part is still being checked. The number of DSs that demand partly lexicalized description is less than 10%. All the rest can be described universally.

3.3. Cross-language asymmetry

Of course, there are a lot of cases of asymmetry between Japanese and other languages of the model, which concern lexicon, voice system, serial verb constructions, copula absence in complement constructions with predicative adjectives, classifiers, or counter words, and a number of other things. The description of these cases is problematic for automation and demands manual work. Different kinds of asymmetry demand different solutions. Due to the lack of space, we cannot provide their detailed description here, and will have to restrict ourselves with a few examples.

Nevertheless, there are many asymmetry cases between the languages that have already been integrated in the model, and the basic principles for dealing with language asymmetry (such as using transformational rules and collocations) are the same for all languages of the model, including Japanese. A detailed description of the methods we use for it is given in [Petrova 2014]. Some instances from Chinese, German and French are suggested in [Manicheva et al. 2012] as well.

As a Japanese instance, let us take lexical asymmetry. There are many concepts in Japanese that are special for Japan, so we do not have SCs for them, like 温泉 [onsen] ‘hot spring’, 炬燵 [kotatsu] ‘Japanese table’, 先輩 [sempai] ‘senior’, and so on. In addition, Japanese abounds with ‘compound’ words, like 陳述書 ‘written declaration’, 魔界 ‘world of spirits’, or 遅咲き ‘late blooming’.

In cases like these, we usually have to add new SCs to the hierarchy, put the required Japanese concepts there and provide their correspondences in other languages of the model. If the equivalent is not a word but a collocation (like ‘hot spring’), we fill the SCs with *terms*—collocations that can be put in particular places of the SH like lexical classes.

The particular cases are concepts composed of antonyms (or somehow opposed notions), or the notions that usually come together: 男女 ‘men and women’, 和戦 ‘war and peace’, 花鳥 ‘flowers and birds‘. Since ‘flower’ and ‘bird’ are definitely two different notions, the corresponding words are “situated” in different SCs, so it is not quite clear where we should place ‘花鳥’. Anyway, in most cases, we have to add such

words to the SH as well, as it facilitates lexical analysis: otherwise, each time the model would have to choose between analyzing the hieroglyph separately, or as a part of a compound word.

4. The Compreno parser and its application

The semantic and syntactic patterns discussed above serve as the basis for the Compreno parser, which includes several other patterns, such as morphology, non-tree links like conjunction, anaphora, and others as well (for detailed analysis, see [Anisimovich et al. 2012]; [Bogdanov, Leontyev 2013]).

Analyzing a sentence, the parser finds a set of syntactic structures that can be matched to it, and then ranges them according to their evaluation. As a result, the parser builds syntactic-semantic structures with the following nodes: SSs and DSs, lexical and semantic classes, semantic and grammatical value, non-tree links—for example, it finds a host for each pronoun, and so on (non-tree links are of great importance for parsing, however, due to the lack of space, we have to omit their description here and refer to the papers mentioned above).

An illustration of the parser's work is figure 1, where the output tree for the English sentence (3) is given:

(3) *I gave a present to my friend.*

Fig. 1. English output tree

In other words, our parser builds a representation of a sentence or a text. Although the current model-based approach demands relatively higher costs, the model shows good evaluation results within the tests such as the "Dialogue evaluation competitions" (http://www.dialog-21.ru/evaluation/), which include a wide range of tasks: morphological analysis, anaphora, entity and fact extraction, machine translation and others (for details, see [Anastasyev et al. 2017]; [Stepanova et al. 2016]; [Bogdanov et al. 2014]; [Zuev et al. 2013]). Moreover, automation methods help to reduce the costs significantly.

As Japanese description is still in progress, we have not made comparative evaluations with other Japanese parsers yet. However, the important idea is that application of the parser to different NLP tasks is to a large extent based on the universal model. For instance, our data extraction mechanism used for English and Russian now relies on the information it gets from the SCs, DSs and tree dependencies—namely, universal objects, which are not language-specific. It means that when we add a new

language to the parser, we still use the same universal structures that we referred to when working with English or Russian. Therefore, most of the IE rules would work for the Japanese IE as well, which helps us to avoid significant work when starting to use the Japanese parser for the tasks like this one.

As stated above, the number of Japanese lexical units added to the system is currently rather modest in comparison with English and Russian. Moreover, so far the syntactic description of Japanese is not complete either. Nevertheless, the Japanese parser functions already and fulfils the semantic and syntactic analysis on limited text collections (the reference to the treebank is given below in section 5). As an instance of Japanese parsing, see example (4) and figure 2 with its output tree (the sentence comes from the open treebank below):

(4) 日本にはたくさんの美しい場所がある。— *There are many lovely places in Japan.*

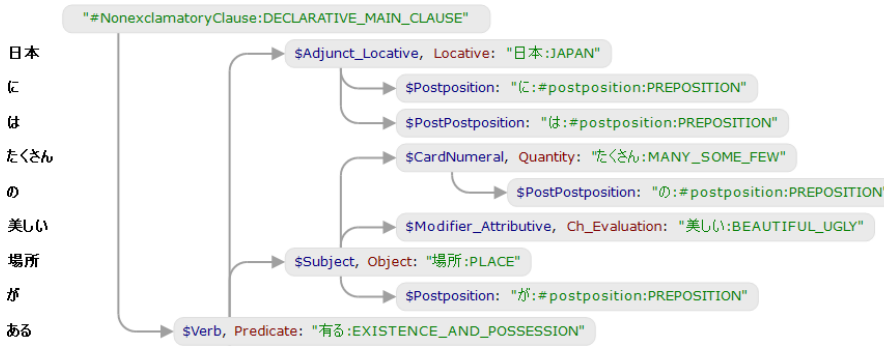


Fig. 2. Japanese output tree

5. Text annotation and Japanese treebank

To evaluate quality change of the parser, we use manually annotated text collections for all languages integrated in the model. Usually, annotation includes DSs and SSs for all constituents.

Our annotation standards have quite a lot in common with the UD principles. Yet, there are significant differences as our annotation is aimed at our project needs and correlates with different opportunities of the model. Unlike the UD, we reconstruct ellipted constituents, which is important for correct information extraction. We treat coordination differently, as in Compreno, all conjuncts are attached to one parent, while in the UD, the first conjunct attaches the other ones. In the UD, punctuators are linked to other constituents, while in Compreno, punctuator is an attribute of a SS, and so on. In general, our annotation demands more competence from the annotator, but gives more precision for our needs.

In future, we plan to open access to some of our annotated corpora, therefore, the opportunity to convert our annotation in the UD standard is in question now.

At the moment, our Japanese treebank consisting of 1,500 sentences is available here: https://github.com/ComprenoData/JapaneseTreebank. The original texts come from the Tatoeba project (https://tatoeba.org/eng), and these are annotated with shallow constituent borders by means of our parser. In addition to the treebank, we suggest the annotation manual at the treebank website, where the annotation syntax and principles are described in greater detail.

6. Conclusion

Integrating Japanese in a formal multilingual model is a challenging task, which faces quite a few difficulties. Nevertheless, the Compreno model proved to be an effective tool for dealing with languages of different groups.

First, the universal SH suits well for word positioning of the lexicon of different languages, including the asymmetry cases. Second, the system of the universal DSs and the inheritance principle allow us to provide each word with all possible semantic links at once purely through the word's positioning in the SH. Third, the architecture of the model reduces significantly the amount of work on the syntactic pattern as well, as the syntactic realizations of the DSs can be introduced high in the SH and be inherited by the SC-descendants.

Though such a model-based approach is rather costly, application of the auxiliary tools and machine learning methods helps to reduce the costs significantly, facilitates and speeds up the description process, and allows us to avoid a number of mistakes inevitable in manual work.

Currently we are continuing to enlarge the Japanese vocabulary added to the SH, progress with the work on the Japanese syntax and start testing the Japanese part of the model on larger text corpora in order to evaluate the current level of the description and to track its progress. At the same time, we plan to focus on the practical application of the Japanese parser and use it for solving different NLP tasks, in particular, for information extraction.

References

1. *Anastasyev D. G., Andrianov A. I., Indenbom E. M.* (2017), Part-of-speech Tagging with Rich Language Description.In Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference "Dialogue 2017", vol. 1, pp. 2–13.
2. *Anisimovich K., Druzhkin K., Minlos F., Petrova M., Selegey V., and Zuev K.* (2012), Syntactic and semantic parser based on ABBYY Compreno linguistic technologies. In Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue", vol. 11, pp. 91–103.
3. *Bogdanov A. V., Dzhumaev S. S., Skorinkin D. A., and Starostin A. S.* (2014), Anaphora analysis based on ABBYY Compreno linguistic technologies. In Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue", pp. 89–101.

4. *Bogdanov A. V., Leontyev A. P.* (2013), Description of the Russian External Possessor Construction in a Natural Language Processing System. In Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue". [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2013] Bekasovo, pp. 110–118.
5. *Goncharova M., Kozlova E., Pasyukov A., Garashchuk R., and Selegey, V.* (2015), Model-based WSA as means of new language integration into a multilingual lexical-semantic database with interlingua. In Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue", vol. 1, pp. 169–182.
6. *Kanayama H., Miyao Y., Tanaka T., Mori S., Asahara M., Uematsu S.* (2015), A draft of universal dependencies for Japanese. In the 21st annual meeting of the Association for Natural Language Processing, pp. 505–508.
7. *Kawahara D., Kurohashi S.* (2006), A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL 2006), pp. 176–183.
8. *Kawahara D., Hayashibe Y., Morita H., Kurohashi S.* (2017), Automatically Acquired Lexical Knowledge Improves Japanese Joint Morphological and Dependency Analysis. In Proceedings of the 15th International Conference on Parsing Technologies, Pisa, pp. 1–10.
9. *Kudo T., Matsumoto Y.* (2002), Japanese dependency analysis using cascaded chunking. In Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002), Vol. 20, pp. 1–7.
10. *Kurohashi S., Nagao M.* (1994), A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. Computational Linguistics, 20(4), pp. 507–534.
11. *Manicheva E., Petrova M., Kozlova E., and Popova T.* (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. Mumbai, pp. 215–229.
12. *Nivre J., de Marneffe M.-C., Ginter F., Goldberg Y., Hajic J., Manning C. D., McDonald R., Petrov S., Pyysalo S., Silveira N., Tsarfaty R., and Zeman D.* (2016), Universal dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari (Conference Chair) et al., editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, may. European Language Resources Association (ELRA), pp. 1659–1666.
13. *Petrova M. A.* (2014), The Compreno Semantic Model: The Universality Problem. In International Journal of Lexicography, Volume 27, Issue 2, pp. 105–129.
14. *Tanaka T., Nagata M.* (2015), Word-based Japanese typed dependency parsing with grammatical function analysis. ACL (2), pp. 237–242.
15. *Tanaka T., Miyao Y., Asahara M., Uematsu S., Kanayama H., Shinsuke M., and Matsumoto Y.* (2016), Universal dependencies for Japanese. In Proceedings of the 10th International Conference on Language Resources and Evaluation. LREC, pp. 1651–1658.

16. *Stepanova M. E., Budnikov E. A., Chelombeeva A. N., Matavina P. V., and Skorinkin D. A.* (2016), Information Extraction Based on Deep Syntactic-Semantic Analysis. In Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue", pp. 721–732.
17. *Uchimoto K., Murata M., Sekine S, and Isahara H.* (2000), Dependency model using posterior context. In Proceedings of the 6th International Workshop on Parsing Technology, pp. 321–322.
18. *Zuyev K. A., Indenbom E. M., and Yudina M. V.* (2013), Statistical machine translation with linguistic language model. In Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue", vol. 2, pp. 175–183.

CORPUS SIZE AND THE ROBUSTNESS OF MEASURES OF CORPUS DISTANCE[1]

Piperski A. Ch. (apiperski@gmail.com)
Russian State University for the Humanities / National Research University Higher School of Economics, Moscow, Russia

This paper studies the impact corpus size has on the robustness of various frequency-based measures of corpus distance (or similarity, respectively), such as Euclidean distance, Manhattan distance, Cosine distance, χ^2, Spearman's ρ, and Simple-Maths Keyword distance. An experiment performed using the British National Corpus shows that Euclidean distance is least influenced by corpus size and thus is best suited for the purpose of comparing corpora.

Keywords: corpus similarity, distance between corpora, evaluation, British National Corpus

РАЗМЕР КОРПУСА И УСТОЙЧИВОСТЬ МЕР РАССТОЯНИЯ МЕЖДУ КОРПУСАМИ

Пиперски А. Ч. (apiperski@gmail.com)
РГГУ / НИУ ВШЭ, Москва, Россия

В статье рассматривается вопрос о том, какое влияние размер корпуса оказывает на устойчивость различных мер сравнения корпусов на основе частотных списков. Для анализа берутся шесть мер: Евклидово расстояние, манхэттенское расстояние, косинусное расстояние, χ^2, ρ Спирмена и сходство по ключевым словам. Эксперимент на материале Британского национального корпуса продемонстрировал, что Евклидово расстояние наименее подвержено влиянию размера корпуса и поэтому лучше всего подходят для сравнения корпусов.

Ключевые слова: сходство корпусов, расстояние между корпусами, эвалюация, Британский национальный корпус

[1] This work is supported by the Russian Science Foundation under grant 17-78-10196.

1. Introduction

The problem of text and corpus similarity is extremely important for Natural Language Processing as well as for corpus linguistics[2]. Measuring similarity (or distance, respectively) is used for information retrieval, text classification, document clustering, machine translation evaluation, and many other applications. A survey of text similarity measures by [Gomaa & Fahmy 2013] includes two types of measures: character-based and term-based measures. Character-based measures treat texts (or text corpora) as sequences of characters which can be transformed into each other by using allowed edit operations, by finding an optimal alignment of two strings, etc. The best-known character-based measure of text similarity is the Levenshtein distance. However, character-based similarity measures imply that similar texts are obtained from each other by some simple operations, which is true in case of shorter texts like misspelled words deriving from the intended correct ones, but it does not conform to our intuition as to how longer texts are produced. For this reason, term-based measures, which can also be called frequency-based measures, are more widely used for measuring similarity between longer texts. To apply these measures, texts are represented as frequency lists, which are then being compared using various ways of measuring distances between vectors, the best-known of them being geometric measures such as Manhattan distance, Euclidean distance, and Cosine simlarity (or distance, respectively), and set-theoretic measures such as Jackard distance.

In corpus linguistics, the problem of text and corpus similarity has gained attention starting with [Kilgarriff 1997] and [Kilgarriff and Rose 1998]. Since then, many measures of corpus similarity have been proposed (see [Kilgarriff 2001, 2009]; [Fothergill et al. 2016]). However, no definitive measure for corpus similarity has yet been found. Specific approaches to computing similarity have been adopted in machine translation evalution, such as BLEU [Papineni 2002], NIST [Doddington 2002], and METEOR [Lavie and Agarwal 2007]. In most other applications that make use of measuring distances between texts, there is no measure that has become a de facto standard, though geometrical measures are generally preferred.

2. Measures of corpus distance

In this paper, I will discuss six measures of distance between corpora. Three of them are based on geometrical notions, namely Euclidean distance, Manhattan distance, and Cosine distance. Two further measures are closely linked to the established statistical procedures; these two measures are χ^2 and Spearman's ρ, which were especially popularized by [Kilgarriff 2001], who showed that they are by far superior to perplexity-based measures. A further measure is Simple-Maths Keyword distance, introduced by [Kilgarriff 2009] and implemented in the Sketch Engine corpus management system ([Kilgarriff et al. 2014]; http://the.sketchengine.co.uk). In this paper, all measures are computed based on the frequencies of 200 most common words

[2] In this paper, I refrain from pursuing the question whether we should using different measures of similarity for individual texts and for collections of texts, i.e., corpora.

in the aggregated frequency distribution of the two corpora being compared (for the first five measures), or on the keyness score of the top 200 keywords in the aggregated keyword list for Simple-Maths Keyword distance.

3. Corpus size and corpus distance

The measures of corpus distance are typically evaluated ussing the Known-Similarity Corpora (KSC) approach [Kilgarriff 1997, 2001]; [Kilgarriff and Rose 1998][3]. A KSC-set is built starting with two corpora X and Y that are deemed to be sufficiently distinct. These original corpora are split into equal-sized chunks that are randomly allocated to new corpora Z_0, Z_1, …, Z_M, each of them consisting of M chunks. Z_0 includes 0 chunks from X and M chunks from Y, Z_1 includes 1 chunk from X and M—1 chunks from Y, Z_2 includes 2 chunks from X and M—2 chunks from Y, etc. The similarity of these corpora is known: for instance, one can assume that Z_3 is closer to Z_5 than Z_2 is to Z_8. Thus, for any $k \leq l < m \leq n$ ($k \neq l$ or $m \neq n$) a good distance measure must satisfy the inequality $d(Z_l, Z_m) < d(Z_k, Z_n)$. One can test whether such an inequality is satisfied for all possible values of k, l, m, and n, and the proportion of inequalities captured correctly indicates how well a distance measure performs. In case of $M = 10$, a total of 660 KSC judgments of the kind need to be tested.

More recent studies have continued this approach, using a wider range of measures and larger amounts of test data [Piperski 2017a, 2017b]. However, no investigations have yet adressed the question of how corpus size influences the robustness of distance measures.

This is a question that plays a significant role in many areas of corpus linguistics. Namely, we can trust a measure of distance only if it yields comparable results when comparing samples from the same populations regardless of sample size or, at least, starting from a certain corpus size. Otherwise, the results might turn out to be untrustworthy6 especially when different-sized corpora are being compared. For this reason, the aim of the present study is to test the robustness of the six measures listed in Section 2 with respect to corpus size. Even though it was shown by [Piperski 2017a] that levels of analysis and segmentation other than individual words, first and foremost character ngrams, are better suited for assessing distance between corpora, in the present study I stick to the word level, since a word is the largest unit that seems to be more or less easily identifiable in a text as well as linguistically significant.

4. Experiment design

For the purpose of the experiment, 200,000-token subcorpora from 11 sources from the British National Corpus (BNC) were taken. The sources are listed in Table 1:

[3] Another approach to this problem is [Forsyth ad Sharoff's 2014] anchor-text method.

Table 1. List of sources

ID	Source	BNC file IDs
art	The Art Newspaper	CKT–CKY, EBS–EBX
bel	The Belfast Telegraph	HJ3–HJ4, K29–K35
bio	The Dictionary of National Biography: Missing persons	GSX–GTH
han	Hansard Extracts	HHV–HHX
ind	The Independent	A1D–A5X
kee	Keesings Contemporary Archives	HKP–HLT
law	The Weekly Law Reports	FBS–FE3
mir	The Daily Mirror	CH1–CH3, CH5–CH7
nsc	New Scientist	ANX, B71–B7N
sco	The Scotsman	K56–K5M
uni	Unigram X	CMW–CN0, CS8–CTV

Obviously, if one is to trust the results of ths study, one must assume that these sources are homogeneous. There is no way of measuring this unless we have a good measure of similarity at our disposal, since homogeneity is closely related to similarity; for this reason, we are forced to take the suitability of the sources for granted.

For each of the sources, 50 random parts having the length of 20,000 tokens[4] (1/10 of the total), 40,000 tokens (2/10), 60,000 tokens (3/10), …, and 180,000 tokens (9/10) were taken, which mimics the approach of Tweedie and Baayen (1998) to measuring lexical diversity. For each pair of sources and for each part size, 50 distances between corresponding random parts were computed, their mean was calculted, and then the 95% confidence interval for the "true" mean of the distance value was estimated using 10,000-sample bootstrapping. For instance, for parts from *The Daily Mirror* and *Unigrams* comprising 100,000 tokens, the 50 Manhattan distances are as follows:

> 0.562; 0.564; 0.567; 0.567; 0.569; 0.569; 0.57; 0.57; 0.571; 0.572; 0.573; 0.573; 0.573; 0.574; 0.576; 0.576; 0.577; 0.578; 0.58; 0.58; 0.58; 0.582; 0.584; 0.585; 0.585; 0.587; 0.592; 0.592; 0.593; 0.598; 0.599; 0.6; 0.606; 0.607; 0.613; 0.614; 0.617; 0.624; 0.625; 0.626; 0.627; 0.628; 0.63; 0.631; 0.632; 0.637; 0.638; 0.64; 0.642; 0.642,

the mean is 0.596, and the 95% confidence interval for the "true" mean is [0.5892; 0.6033]. This confidence interval can be further compared to the best estimate of the distance between the two sources, namely the distance between the two 200,000-token corpora—in this case, 0.5895. This distance falls withing the estimated confidence interval, which is an indicator of the fact that this measure is robust with respect to corpus size, because it yields a good estimate of corpus distance even with a relatively small corpus size. If the best estimate (i.e., the estimate based on 200,000-token portions) falls within the confidence interval for a certain pair of sources and for

[4] All manipulations with the BNC were performed using Python 3.6, and, more specifically, the BNCCorpusReader class from NLTK [Bird et al. 2009]. Punctuation marks were treated as separate tokens, and word processing was case-sensitive.

a certain corpus size, the measure gets 1 point; otherwise, it gets 0 points. Because we have 11 sources, the number of pairs is $11 \times 10 / 2 = 55$; for each pair, we work with 9 corpus sizes, which makes a total of $55 \times 9 = 495$ test cases for each of the six measures. The winning measure is the one that gets the most out of 495 possible points.

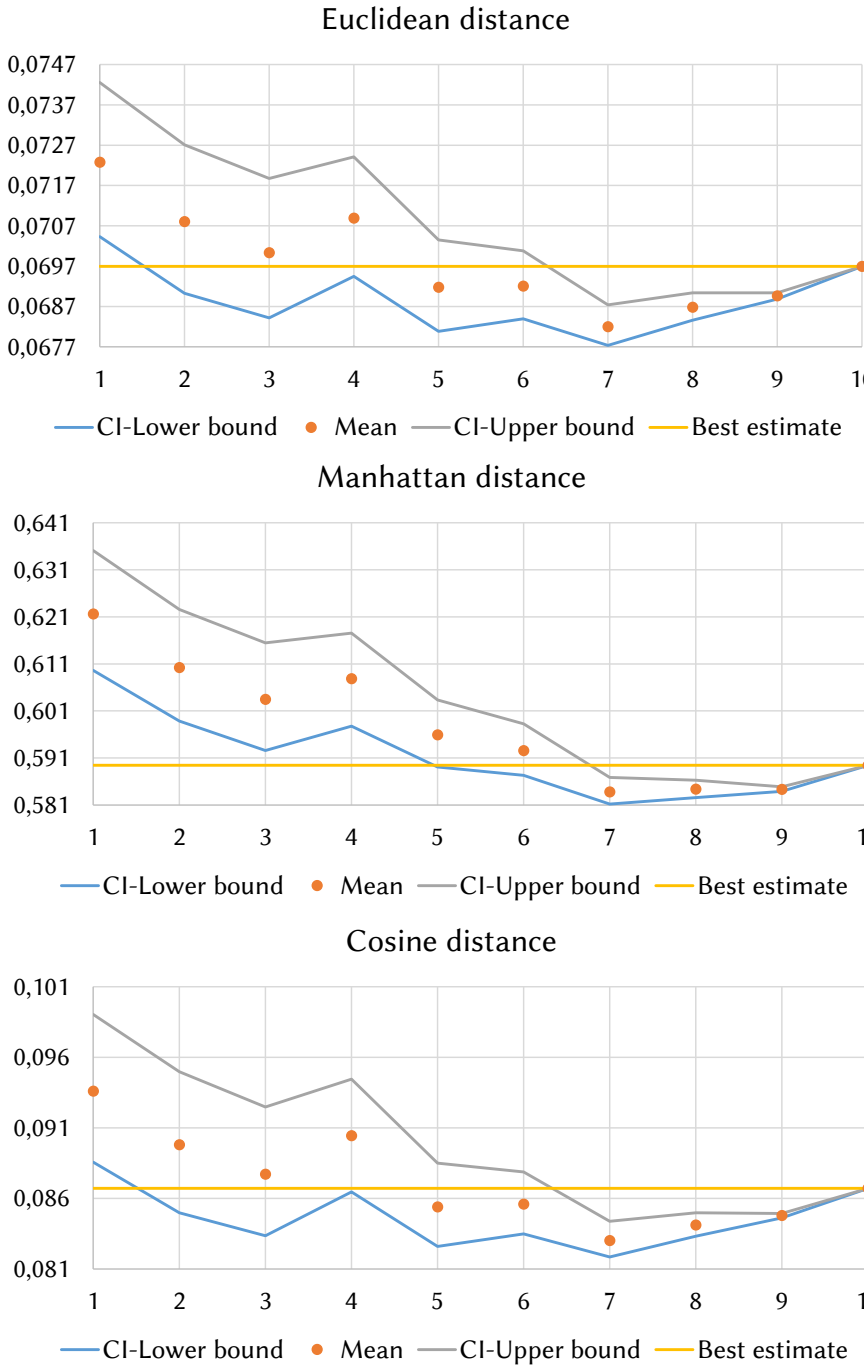
Obviously, when we measure a distance between two corpora, we never now whether the resulting distance falls close to the "true" dustance based on the whole populations. However, if we know in advance that a measure often comes close the best estimate we have, this might speak in favor of this measure.

5. Results

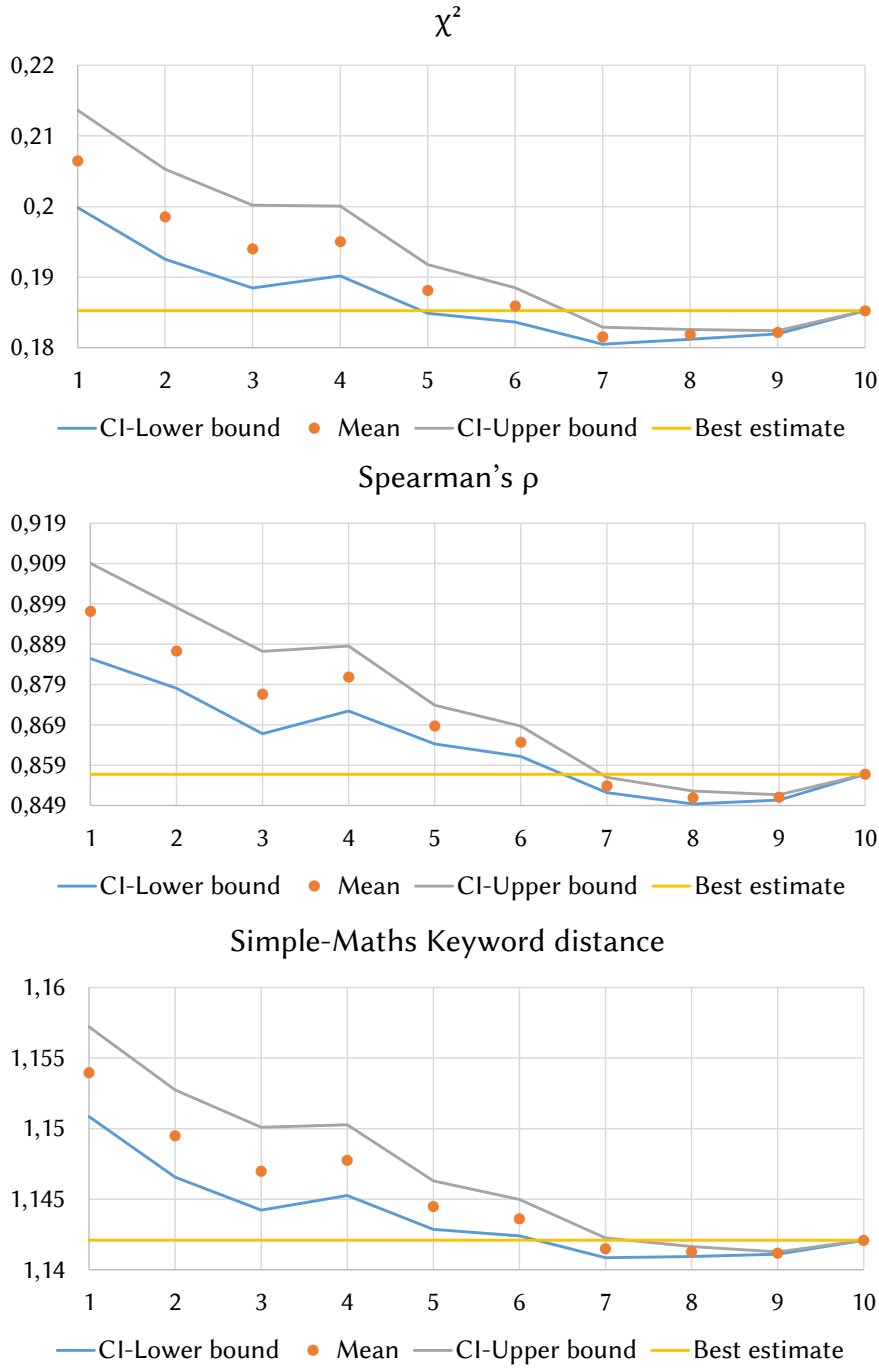
As an illustration, the result for one pair of corpora is shown in Figures 1 to 6 below. They visualize the comparison of *The Daily Mirror* with *Unigram X*. One can see from Figure 1 that the best estimate for the two sources as a whole falls within the bounds of the confidence interval in 5 cases out of 9 for Euclidean distance, because the orange line lies between the grey and the blue line for $x = 2, 3, 4, 5, 6$ (correspoding to 40,000-token, 60,000-token, 80,000-token, 100,000-token, and 120,000-token corpora). It also falls within the bounds of the confidence interval in 2 cases out of 9 for Manhattan distance, in 5 cases out of 9 for Cosine distance, in 2 cases out of 9 for χ^2, in 0 cases out of 9 for Spearman's ρ, and in 1 case out of 9 for Simple-Maths Keyword distance. Thus, in this case the two most robust measures are Euclidean distance and Cosine distance, whereas Spearman's ρ is the least robust one.

Interestingly, in all cases the distances obtained for smaller corpora are larger than the best estimate, which also holds true for other pairs of sources. The amount of variation in estimates for smaller corpora is not surprising, because larger parts must overlap with each other, whereas smaller parts do not necessarily do so. The form of the graphs is similar in all six cases (a fall, then a rise at 4/10 of the corpus, then a further fall followed by a rise), but this is an artifact of random sampling from *The Daily Mirror* and *Unigram X*; for another pair of sources, the graphs need not look the same.

The total counts of the best estimates falling within the counfidence intervals for smaller parts of corpora based on all 55 pairs of sources are presented in Table 2:



Figures 1–3. Robustness of the six distance measures as comapred using *The Daily Mirror* and *Unigram X*



Figures 4–6. Robustness of the six distance measures as comapred using *The Daily Mirror* and *Unigram X*

Table 2. Overall robustness of the six distance measures

Distance measure	Score	Percentage
Euclidean	91	18%
Manhattan	55	11%
Cosine	84	17%
χ^2	50	10%
Spearman's ρ	43	9%
Simple-Maths Keywords	42	8%

This table shows that Euclidean distance and Cosine distance are the most robust measures with respect to corpus size, whereas other measures, including both statistical measures and the keyword-based measure, are less trustworthy. This also conforms to the findings by [Piperski 2017b], who showed that geometrical measures of corpus distance perform best when assessed with the Know-Similarity Corpora approach.

6. Stability of the confidence interval

A further question arises from the fact that the evaluation technique presented above can be easily tricked. Namely, if a measure provides a wide confidence interval for some smaller corpus size, it is likely that the "true" estimate will fall within this interval. This suggests an additional requirement on the winning measure: it must not inflate the confidence interval for smaller sample sizes, i.e. its estimates for the same corpus size must not be too different from each other. The problem is that the width of the confidence interval is hard to compare across different measures. We cannot just express it as a percentage of the absolute value of the best estimate, because adding a constant to the distance would not change the measure as such, but it would change this percentage; for instance, the Simple-Maths Keyword distance as implemented in SketchEngine has a minimum value of 1, and if we were to subtract 1 from it, we would assess the relative width of the confidence interval differently.

To counter these difficulties, I propose two measures of stability of the confidence interval. First, as already mentioned in Section 5, it is evident that the variation of distance estimates between smaller corpora must be larger that the variation of distance estimates between larger corpora, simply because larger corpora drawn from the same 200,000-word population must necessarily overlap. We expect the confidence interval for 180,000-word corpora to be smaller than the confidence interval for 20,000-word corpora, and we can calculate how many times larger is the confidence interval for the mean for the smallest corpus size (20,000 tokens) as compared to the largest corpus size (180,000 tokens, since we do not have a confidence interval for 200,000-token corpus, but only a single estimate).

For example, if we apply Manhattan distance to the corpora sampled from *The Daily Mirror* and *Unigram X*, the confidence interval is [0.6097, 0.6347] for 20,000-token corpora and [0.5839, 0.5849] for 180,000-token corpora. This means that making the corpus 9 times smaller increases the confidence interval by 25 times. This value can be computed for each measure for all 55 pairs of sources. The results are summarized in Table 3.

Table 3. The increase of the width of the confidence interval from 180,000-token to 20,000-token corpora

Distance measure	Mean	Median
Euclidean	17.8	16.8
Manhattan	19.4	18.3
Cosine	21.7	19.4
χ^2	25.1	21.5
Spearman's ρ	17.0	15.4
Simple-Maths Keywords	22.7	20.8

Table 3 shows that the two best measures in this respect are Euclidean distance and Spearman's ρ. Cosine distance has performed well during the first test, but it might be due to the fact that it is likely to inflate the confidence interval.

Second, we also must check whether a distance measure is biased in some direction with respect to corpus size. Even if a measure has a relatively stable confidence interval, it may be the case that this interval is gradually drifting away from the best estimate the smaller our corpus becomes. This means that if we take one step further towards a smaller corpus, we must accept it that the confidence interval becomes larger, but we cannot tolerate if it steadily shifts in one direction. In the graphs above, there is a somewhat unsatisfying general upwards trend when looking from right to left. In order to quantify this trend, I propose the following way of computing an instability score: for $1 \leq n \leq 8$, if a distance measured for a pair of corpora containing $n \times 20{,}000$ tokens is larger than the upper bound of the confidence interval for the mean distance for $(n + 1) \times 20{,}000$ tokens, the measure is given $n / (n + 1)$ points[5]; if, on the contrary, a distance is smaller than the lower bound of the confidence interval, the measure loses $n / (n + 1)$ points. A good measure will behave symmetrically, i.e., it will receive approximately the same amount of points as it will lose, making the result close to 0.

In the worst-case scenario, a measure may receive a score whose absolute value is $(8/9+7/8+6/7+5/6+4/5+3/4+2/3+1/2) \times 50 \times 55 \approx 16{,}970$ (the value would be negative the distances are decreasing with decreasing corpus size, and positive otherwise).

Table 4. Instability scores

Distance measure	Instability score
Euclidean	2,072.5
Manhattan	4,365.9
Cosine	2,345.4
χ^2	5,462.2
Spearman's ρ	4,990.3
Simple-Maths Keywords	6,453.3

[5] Since corpora of $n \times 20{,}000$ tokens are in generally more similar to $(n + 1) \times 20{,}000$-token corpora, falling outside the confidence interval must cost more for larger values of n. However, the proposed cost of $n / (n + 1)$ was selected *ad hoc* and has no external justification.

Table 4 shows that all measures have a positive instability score, i.e. they tend to yield larger distances when corpus size decreases. The two measures with the smallest value of instability score are Euclidean distance and Cosine distance.

7. Conclusion

This paper presents the results of testing the frequency-based measures of corpus distance the for robustness with respect to corpus size. In all three experiments (Tables 2 to 4), Euclidean distance was among the two best measures, which leads us to a conclusion that it is actually the measure that is most robust to corpus size among the six measures evaluated. Further possible directions of study include evaluating robustness of distance measures when measured for corpora of different sizes as well as taking into consideration languages other than British English only.

References

1. *Bird S., Klein E., Loper E.* (2009), Natural Language Processing with Python, Cambridge (Mass.), O'Reilly Media.
2. *Gomaa W. H., Fahmy A. A.* (2013), A Survey of Text Similarity Approaches, International Journal of Computer Applications, 68(13), April, pp. 13–18.
3. *Doddington G.* (2002), Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics, In Proceedings of the Second International Conference on Human Language Technology Research, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc, pp. 138–145.
4. *Forsyth, R. S., Sharoff S.* (2014), Document dissimilarity within and across languages: A benchmarking study, Literary and Linguistic Computing, 29:1, pp. 6–22.
5. *Fothergill R., Cook P., Baldwin T.* (2016), Evaluating a topic modelling approach to measuring corpus similarity, In Proceedings of the Tenth International Conference on Language Re-sources and Evaluation (LREC 2016), Portorož, Slovenia, pp. 273–279.
6. *Kilgarriff A.* (1997), Using word frequency lists to measure corpus homogeneity and similarity between corpora, http://aclweb.org/anthology/W97-0122.
7. *Kilgarriff A.* (2001), Comparing corpora, International Journal of Corpus Linguistics, 6(1), pp. 97–133.
8. *Kilgarriff A.* (2009), Simple maths for keywords, In Proceedings of Corpus Linguistics Conference CL2009, University of Liverpool, UK, July 2009.
9. *Kilgarriff A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Rychlý P., Suchomel V.* (2014), The Sketch Engine: ten years on, Lexicography, 1:1, pp. 7–36.
10. *Kilgarriff A., Rose T.* (1998), Measures for corpus similarity and homogeneity, http://aclweb.org/anthology/W98-1506.
11. *Lavie A., Agarwal A.* (2007), Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments, In Proceedings of the Second Workshop on Statistical Machine Translation, pp. 228–231.

12. *Papineni K., Roukos S., Ward T., Zhu W.-J.* (2002). BLEU: A Method for Auto-matic Evaluation of Machine Translation, In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318.
13. *Piperski A.* (2017a), Sravnenie korpusov meroj χ^2: simvoly, slova, lemmy ili časterečnye pomety? [Comparing corpora with χ^2: characters, words, lemmata, or PoS tags?], In Korpusnaja lingvistika–2017 [Corpus Linguistics–2017], Saint Petersburg, Saint Petersburg State University, pp. 282–286.
14. *Piperski A.* (2017b), Izmerenie rasstojanij mezhdu korpusami [Measuring distances between corpora], course given at Tampere Summer School in Multilingual Corpora, Tampere, Finland, 28 August–1 September 2017.
15. *Tweedie F. J., Baayen R. H.* (1998), How Variable May a Constant be? Measures of Lexical Richness in Perspective, Computers and the Humanities, 32(5), pp. 323–352.

«А У НАС В КВАРТИРЕ ГАЗ! А У ВАС?»: КОНСТРУКЦИИ С СОЮЗОМ *А* ПО ДАННЫМ ПРОСОДИЧЕСКИ РАЗМЕЧЕННОГО КОРПУСА

Подлесская В. И. (vi_podlesskaya@il-rggu.ru)

Российский государственный гуманитарный университет,
Российская академия народного хозяйства
и государственной службы;
Москва, Россия

Ключевые слова: сложное предложение, русский язык, корпус, устная речь, просодия

"A U NAS V KVARTIRE GAZ. A U VAS?": THE RUSSIAN CONJUNCTION *A* VIEWED THROUGH THE PRISM OF PROSODICALLY ANNOTATED CORPUS DATA

Podlesskaya V. I. (vi_podlesskaya@il-rggu.ru)

Russian State University for the Humanities,
Russian Academy of National Economy and Public Administration;
Moscow, Russia

The paper focuses on Russian constructions with clauses (or VPs) combined by means of the discourse marker *A*, that behaves as a conjunction or as a particle in different contexts. Prosodically, the construction may come up in two forms: (a) as a single illocution with the first clause pronounced with a rising pitch that projects discourse continuation, and (b) as two separate illocutions with the first clause pronounced with a falling pitch that projects no continuation. Basing on the data from the Prosodically Annotated Corpus of Spoken Russian, prosody and grammar of (a) and (b) were analyzed qualitatively and quantitatively. Type (b) appeared to be as frequent as type (a) and systematically favored in pragmatically marked contexts.

Key words: clause combining, Russian, corpus, natural discourse, prosody

1. Постановка вопроса

В работе на материале просодически размеченного корпуса устных личных рассказов исследуются сегментные и супрасегментные свойства дискурсивного маркера А — союза и «смежной» с союзом частицы[1]. Синтаксису и особенно, семантике этого маркера посвящена огромная литература, включающая, в частности, ставшие уже классическими сочинения [Зализняк, Микаэлян 2005], [Йокояма 1990], [Крейдлин, Падучева 1974а,б], [Левин 1970], [Санников 2008: 108–382], [Урысон 2006, 2011], [Янко 1990, Malchukov 2004] и ряд других. Систематическое исследование семантики и дистрибуции этой единицы на корпусном материале содержится в статье [Зализняк, Микаэлян 2018], там же приводится обширная литература. Вместе с тем, просодия конструкций с А остается в значительной степени белым пятном.

Ядерные употребления союза А — биклаузальные конструкции, в которых с помощью союза сопоставлены два положения дел, у которых обнаруживаются сходства и различия по тому или иному параметру. Так в примере (1)

(1) (контекст: *Где это видано? Где это слыхано?)*
Дедушка едет, а мальчик идет! (С. Маршак)

говорящий противопоставляет два одновременно наблюдаемых способа перемещения двух действующих лиц. Такое употребление маркера можно условно назвать «двухместным». Синтаксическая интеграция клауз в таких конструкциях проявляется, в частности, в том, что они попадают в сферу действия правил эллипсиса и анафорической замены, ср. *Я еду$_i$ на поезде, а ты — Ø$_i$ — на машине*. Синтаксическая интеграция клауз в таких конструкциях, как правило, подкрепляется просодической интеграцией: так, устная версия примера (1), в общем случае, реализуется с подъемом тона в главном фразовом акценте первого компонента, который располагается на ударном слоге слова *едет*; этот подъем является стандартным просодическим средством оформления дискурсивной незавершенности в русском языке. В письменной речи просодическая незавершенность конвенционально нотируется с помощью пунктуации, ср. запятую перед А в примере (1).

Наряду с двухместным употреблением в текстах, особенно разговорных, широко представлено употребление, которое можно назвать «одноместным»[2]. Здесь вводится изолированное положение дел, и функция маркера А «привязать» это положение дел к широкому, часто даже экстралингвистическому контексту и показать, что какой-то параметр вводимого положения дел

[1] Исследование выполнено при поддержке РНФ, грант №17-18-01184.

[2] Анна А. Зализняк и И. Микаэлян (2018) используют в близком значении термин «инициальное А», однако трактуют его расширительно, включая сюда и А в функции междометия, в том числе, сигнализирующего «вынужденное согласие», ср. *А! Черт с ним! Как будет, так будет».* Между тем, междометийное употребление демонстрирует принципиально иные просодические свойства — А-частица употребляется строго атонически, тогда как А-междометие, как правило, употребляется просодически автономно, становясь акцентоносителем. Такие употребления в настоящей статье не рассматриваются.

не выводится непосредственно из текущего информационного фона, например, объявить о смене дискурсивной темы (ср. термин «А поворота повествования», введенный [Е. В. Урысон 2006, 2011]). Одноместные употребления особенно характерны для диалогического, а не для нарративного режима, ср. эмфатически нагруженные вопрос (2) и жалобу (3):

(2) *А куда это ты собрался?*

(3) *А Петька дерется!*

Маркер А в одноместном употреблении в русистике обычно квалифицируется как частица. Для частицы А характерна позиция абсолютного начала иллокуции, в том числе, абсолютного начала диалогической реплики. Фрагмент перед ней, в общем случае, может иметь любую иллокутивную силу, например, вопрос, ср. *Ты почему так поздно встал? — А мне сегодня ко второму уроку.* Если фрагмент, предшествующий одноместному А, имеет иллокутивную силу сообщения, то его иллокутивная независимость и просодическая завершенность сигнализируется прототипически в помощью падения тона в главном фразовом акценте, ср. естественное произнесение с падением тона в главном фразовом акценте первой реплики на ударном слоге глагола *устал*: *Что-то я устал. — А чем ты занимался?* В письменной речи такого рода просодическая завершенность конвенционально нотируется с помощью точки и началом следующего фрагмента с заглавной буквы.

Характерной для частицы А является и позиция начала прямой цитаты, следующей за авторской ремаркой, ср. *Ваня тут же пожаловался: «А Петька дерется!»* Препозитивная авторская ремарка, как и независимое сообщение, обычно реализуется с падением тона в главном фразовом акценте. Это создает уникальную дискурсивную коллизию: просодическая реализация не проецирует продолжения, но в то же время, наличие предиката речи с незаполненной валентностью, напротив, предполагает дальнейшее развертывание дискурса. Такой конфликт просодии и лексико-грамматической формы конвенционально нотируется в письменной речи с помощью двоеточия.

Разграничить А-союз и А-частицу в живой речи удается далеко не всегда, ср. следующее показательное свидетельство из [Грамматики-80, 1982]:

> § 1699. Многие частицы по своему значению и по своим синтаксическим функциям не противостоят резко словам других классов — союзам, вводным словам, междометиям, наречиям, а совмещают в себе качества частицы и слова одного из этих классов. ... Ч а с т и ц ы - с о ю з ы совмещают разные модальные значения со значениями связующих слов. ... Частицы *а*, *и* выражают собственно связь, соединенность: — *Да что же это такое*! — *вскричал я*. — *А то такое, что и не знаю, что с ней делать* (Дост.); — *И чудной ты*! — *проговорил вдругЕгоркин*. — *Чем чудной*? — *А всем*! (Станюк.);

Важным тестом, помогающим разграничить А-союз и А-частицу, является тест на синтаксическую подчинимость: конструкции с А-частицей — но не с А-союзом! — относятся к разряду так называемых main-clause phenomena «явлений главного предложения», или, согласно важному уточнению, сделанному Е. В. Падучевой [1996a:299], «явлений предложения, которому соответствует отдельный речевой акт» (ср. также близкий термин root phenomena, [Lobke et al. 2012] и обширную библиографию там же). Это проявляется, в частности, в том, что конструкция с А-союзом может быть подчинена внешнему хозяину, а конструкция с А-частицей — нет, ср. допустимость подчинения в формате косвенной речи в (1а) и недопустимость в (3а):

(1а) *Народ кричал, что дедушка едет, а мальчик идет*

(3а) **Ваня пожаловался, что а Петька дерется!*

Однако даже применение такого мощного теста наталкивается на сопротивление эмпирического материала. Представим себе, что предложение (1) с А-союзом произносится не с подъемом тона на *едет*, а с падением. Зафиксируем это вполне допустимое произнесение с помощью точки перед А:

(1б) *Дедушка едет. А мальчик идет.*

Фактически здесь произошла парцелляция второго компонента конструкции с изменением просодической реализации при сохранении сегментного состава. Передать эту просодическую «точку» перед А в формате косвенной речи уже не удастся. Еще заметнее станет неподчинимость парцеллированного фрагмента конструкции, если мы распределим ее между двумя говорящими:

(1в) А: *Дедушка едет.*
Б: *А мальчик идет.*

Процитировать реплику говорящего Б в формате косвенной речи совершенно невозможно: ** Б говорит, что а мальчик идет*. При этом исходная «двухместная» семантика союза полностью сохранена!

Уже этот простой пример показывает, что коммуникативно-просодическая организация конструкций с маркером А должна составить — наряду с их грамматическими и семантическими свойствами — существенную часть «портрета» этих конструкций. Данная работа представляет собой попытку продвинуться в решении этой задачи с опорой на корпусные данные живой речи. Попробуем ответить на следующие исследовательские вопросы:

- Каков арсенал просодических конфигураций, используемых в составе конструкций с А?
- из этих конфигураций могут использоваться в качестве сигналов
- автономности\неавтономности клаузы в разворачивающейся структуре дискурса?
- Как эти просодические сигналы могут согласовываться (или не согласовываться!) с лексико-грамматическими?
- Отличается ли употребление конструкций с А в устном и письменном дискурсе?

Я использую материал корпуса «Веселые истории из жизни» электронной коллекции «Рассказы о сновидениях и другие корпуса звучащей речи» [Spokencorpora 2013]. Корпус содержит 40 устных монологов (аудиофайлы с синхронизированными просодически размеченными транскриптами, респонденты от 18 до 60 лет, около 10 000 словоупотреблений), плюс письменные версии этих же рассказов, самостоятельно записанные авторами спустя несколько дней после записи устной версии (около 7000 словоупотреблений). Исследуемый корпус предоставляет уникальный материал для сравнения дискурсивных стратегий — в том числе, стратегий предикативного сочинения — в устной и письменной речи именно потому, что включает устные и письменные версии рассказов одного и того же говорящего, основанные на одном и том же сюжете. Были проанализированы все вхождения маркера А, обнаруженные в корпусе: 93 вхождения в устном подкорпусе (далее в тексте — примеры с индексом FS-Sp) и 76 эпизодов в письменном подкорпусе (примеры с индексом FS-Wr). Дальнейшее изложение будет строиться следующим образом: в разделе 2 будет описана просодия и грамматика конструкций с А в устной речи, а в разделе 3 будут привлечены данные письменного подкорпуса, а также — для сравнения — корпусные данные об употреблении семантически и структурно близкого союза НО. В завершении раздела 3 будут подведены количественные и качественные итоги исследования.

2. Употребления дискурсивного маркера *А* в устном подкорпусе «Веселых историй»

Первое важное корпусное свидетельство состоит в том, что конструкции с маркером А с просодической завершенностью фрагмента перед А (условно — «частицы»), встречаются едва ли не чаще, чем конструкции с просодической незавершенностью перед А (условно — «союзы»). Из общего числа конструкций в устном подкорпусе (93) обнаруживается 35 экземпляров с просодической незавершенностью, 41 — с просодической завершенностью и еще 17, в которых это противопоставление либо нейтрализовано, либо его не удается однозначно интерпретировать. Рассмотрим эти три класса случаев более подробно.

2.1. Просодческая незавершенность перед *А*

Как и следовало ожидать, в нашем материале широко представлены прототипические бинарные конструкции с А, компоненты которых иллокутивно однородны, т. е. имеют одинаковую иллокутивную силу (преимущественно, сообщение, так как мы имеем дело с коллекцией нарративов), и первый компонент которых реализуется как просодически незавершенный. Дефолтный способ маркирования просодической незавершенности в таких конструкциях «прототипический русский подъем» с падением на заударных, если они есть [Янко 2008:31], т. е. по типу ИК-3 в терминологии интонационных конструкций [Брызгунова 1982а], ср. подъем на слове *главном* в строке 6 в (4):

(4) FS_37-f_Sp[3]

6. ··· А поскольку-у ·· \ректорат ·· находится в /гл<u>а</u>вном здании,
7. ··· а \физфак и \кафедра в то время ··· располагалась в том числе ··· в /**отд<u>е</u>льном** корпусе,

В единичных случаях незавершенность перед А маркируется конфигурацией типа ИК-4 — падением на ударном слоге с последующим подъемом на заударных или непосредственно на ударном слоге, если заударных нет. Эта конфигурация в русском языке, по наблюдениям Т. Е. Янко [2008: 33, 200–225], связана со значением «рассказа по порядку», а также с сопоставлением и противопоставлением, ср. движение тона на слове *кирпич* в (5):

(5) FS_02-f_Sp

17. ··· ↑\/–Заезж<u>а-а</u>ем мы значит,
18. ·· под \/кирп<u>и</u>ч,
19. ·· а /впереди едет \маш<u>и</u>на,
20. и стоят там \милицион<u>е</u>ры,

Возможна конфигурация с так называемым «нефинальным» падением, [Кибрик, Подлесская (ред.):152–155], т. е. падением не в самый низкий для данного говорящего уровень. Одна из функций нефинального падения как маркера дискурсивной незавершенности — показать, что данное движение тона адаптируется к подъему тона в следующей коммуникативно-просодической составляющей, чтобы интегрировать текущую составляющую с последующей. Именно такая функция в (6) у падения на слове *одной* в строке 25 — оно адаптировано к подъему в главном акценте следующей строки, на слове *другой*. Подъем на слове *другой* вызван внешними дискурсивными причинами — он маркирует незаконченное перечисление. Благодаря адаптивному падению в строке 25, строки 25–26 интегрируются в качестве уточнения к строке 24 и образуют с ней единый элемент в цепочке перечисления (другие элементы — строка 27 и строки 28–29):

(6) FS_37-f_Sp

24. ···· /Вот через такую \лужу нам хотелось /перебр<u>а</u>ться,
25. ему с \одн<u>о</u>й стороны,
26. а мне с /–друг<u>о</u>й →стороны,,,
27. ··· оба /–спеш<u>и</u>ли,,,
28. ··· вблизи \н<u>е</u> было /–в<u>и</u>дно,
29. ·· где-е <ем=> ‖ где бы эту /лужу можно было \обойт<u>и</u>.

[3] Об используемой системе дискурсивной транскрипции см. Кибрик, Подлесская (ред.) 2009, Spokencorpora 2013. Индекс примера содержит отсылку к ярлыку текста в составе корпуса.

Пример (6) иллюстрирует еще одну важную особенность конструкций с просодической незавершенностью перед А: существуют лексические, грамматические и дискурсивные факторы, которые устойчиво коррелируют с просодической незавершенностью. Одним из таких факторов является наличие эллипсиса и анафорических замен, (6), (7):

(7) FS_37-f_Sp

40. вот-т з= || как бы /люди заняты вроде /своим делом,
41. а вроде \нет.

Другим фактором, способствующим просодической незавершенности, являются устойчивые лексико-грамматические конструкции, первая часть которых включает проектор, предсказывающий появление второй — *один.. другой\ второй*, (6); *пустячок, а приятно*, (8); *не Х, а У*, (10), и т. п.:

(8) FS_35-m_Sp

32. ·· ну в общем как бы-ы ·· /пустячок,
33. ·· а \приятно.

(9) FS_18-f_Sp

68. но-о ” не /возмущались,
69. а наоборот \/радовались,
70. ·· за \студентов,

Просодической незавершенности способствуют некоторые конфигурации дискурсивной структуры. В частности, между компонентами, связанными с помощью маркера А двухместным отношением, может иметься вставка, уточняющая значение первого компонента. Эта вставка часто произносится как парентеза — в более узком частотном диапазоне, со сниженным уровнем громкости — и может завершаться нисходящим акцентом, но семантически сферой действия просодической незавершенности является не материал вставки, а именно фрагмент, вводимый маркером А. Так, в следующем примере просодическим проектором является восходящий акцент в строке 39, именно он предсказывает появление строк 41–42, связанных отношением следования со строкой 39, а строка 40 вводит фоновую информацию и просодически реализуется как парентеза:

(10) FS_02-f_Sp

38. ··(0.40) И тут он значит /оборачивается,
39. смотрит на одну из кадок с этими /цветами,
40. (а там \кактусов было много,)
41. ···(0.52) и /говорит,
42. что мол одного кактуса не \хватает.

Сходным образом, просодическая незавершенность в строке 50 (восходящий акцент на слове *аэропорт*) «разрешается» в строках 53–54, связанных с 50 отношением следования, а строка 51 — с А — является парентетической вставкой:

(11) FS_05-m_Sp

50. ··· /И на следующий /день ··· приходит он в /аэропорт,
51. ···· ээ (а /там-м /аэропорт это такое \поле,)
52. ··· \вот,
53. ·· /смотрит,
54. /–самолёт его подогнали,,,

Наконец, просодической незавершенности могут способствовать и семантические факторы. Так, один из распространенных семантических классов конструкций с А вводит во второй части некоторое положение дел, которое обнаруживается субъектом, а в первой части — обстоятельство обнаружения или действие, в результате которого это положение дел обнаруживается. В таких конструкциях первый компонент устойчиво реализуется как просодически незавершенный:

(12) FS_04-f_Sp

27. моя /подружка —
28. /Кристина,
29. — .. {СМЕХ } /смотрит,
30. а там у него \цветы всякие в горшках,

(13) FS_23-f_Sp

112. ·· /–возвращаюсь,
113. ···· {ЧМОКАНЬЕ} ··· а он ей на-а подушку —
114. ·· под /–**покрыва-ало**!,
115. — ··· \накакал,

2.2. Просодческая завершенность перед *А*

Просодическая завершенность перед А стандартно маркируется падением тона в главном фразовом акценте первого компонента по типу ИК-1 в терминологии интонационных конструкций [Брызгунова 1982, Янко 2008]. Среди конструкций с просодически завершенным первым компонентом имеются прототипические биклаузальные иллокутивно однородные конструкции, в которых А можно было бы квалифицировать как семантически двухместный союз, если бы не просодическая дезинтеграция второго компонента (ср. иллюстративный пример (1б) выше). Таковы употребления А в строках 56 и 57 следующего примера:

(14) FS_19-f _Sp

54. и перед нами стоит абсолютно голый \мужчи́на.
55. ···· И \смо́трит на нас.
56. ···· А я смотрю на \него́.
57. А он на \на́с смотрит.

Более заметный сдвиг в сторону частицы по шкале «двухместный союз vs. одноместная частица» наблюдается в тех случаях, когда компонентами семантически двухместного маркера являются не единичные клаузы, а более протяженные фрагменты. Так, в следующем примере оба вхождения А (строки 8 и 11) следуют за просодически завершенными фрагментами, и оба маркируют смену микроэпизодов, связанную со сменой протагониста:

(15) FS_03-m_Sp

7. ·· И-и ····{ЦОКАНЬЕ} значит-т ·· прошу \щи́.
8. ···’А тётка —
9. ·· ээ ···которая ··· эти щи самые ··· /кладёт,
10. — ··· не к= || не кладёт мне в них \смета́ну.
11. А я || ·· я с= || см= || /смотрю́,
12. у них там ·· рядом —
13. ··(0.40) на \сто́йке,
14. — стоит \смета́на,
15. ··(0.34) и явно её ··(0.25) нужно класть в \щи́.

Сходным образом, А в следующем примере маркирует переключение условий при сохранении протагониста:

(16) FS_19-f _Sp

76. ··· Каждый раз когда кто-то /проходи́л ·· мимо двери,
77. если эт= || ээ если в компании проходящих б-был какой-нибудь / мужчи́на,
78. то он не \открыва́л дверь.
79. ·· А если шли только ·· особи /же́нского пола,
80. он значит ·· тут как \/ту́т,

Дальнейшее продвижение в сторону частицы обнаруживается в тех случаях, где компоненты двухместного отношения входят в реплики разных говорящих, ср. иллюстративный пример (1в) выше и (17) — «она землячка, а я живу в одном городе»:

(17) FS_13-f _Sp

54. и он \говори́т:
55. ···«И ещё ·· к тому же ·· она землячка \Медве́дева!»

56. ·· Ну я говорю
57. «/Да?
58. И \/что?
59. ’А **я** с ним в одном \городе живу.»

Наконец, одноместному употреблению А, которое естественно квалифицировать как частицу, в нашем массиве примеров всегда сопутствует просодическая завершенность предшествующего компонента. Такие употребления, как уже говорилось выше, свойственны диалогическому режиму, в то время как исследуемый корпус составляют нарративы. Поэтому неудивительно, что одноместные употребления А обнаруживаются в нашем материале в составе прямых цитат — почти четверть (9 из 41 эпизода просодической завершенности перед А) составляют случаи, где директивная, пример (18), или вопросительная, пример (19), реплика следует в формате прямой речи после авторской ремарки. Ремарка же стандартным образом реализуется с падением тона в главном акценте[4]:

(18) FS_03-m_Sp

20. ··(0.40) Я ··(0.39) и \говорю этой тётке:
21. ···(0.50) «А-а ···(0.65) положите мне э-э ···(0.52) \сметаны в щи¡».

(19) FS_05-m_Sp

45. ··· и тут папа \думает:
46. «А’ || а’ || а собственно ·· ээ так ли велика разница между шестым и \шестнадцатым?»

2.3. Спорные случаи

В исследованном материале употребление А не исчерпывается перечисленными выше очевидными случаями просодической завершенности и просодической незавершенности.

Первую группу проблемных случаев составляют контексты, в которых фрагмент, предшествующий А, маркирован акцентом типа ИК-6 по Е. А. Брызгуновой — с подъемом на ударном слоге, за которым не следует падения на заударных. Эта конфигурация, которую условно можно назвать интонацией многоточия, особенно при растянутом ударном слоге, выражает, согласно Т. Е. Янко, значение имитации ментальной деятельности (припоминание, недоумение), однако при этом широко используется и для выражения незавершенности при описании череды событий, открытого списка [Янко 2008: 109–113, 166–167]. Получая на вход фрагмент, оформленный таким образом, слушающий допускает, что возможно продолжение,

[4] В живой речи препозитивная авторская ремарка может произноситься и атонически, образуя, фактически, единую иллокуцию с цитируемым фрагментом. В этом случае можно считать, что левая граница фрагмента с одноместным А совпадает с левой границей авторской ремарки.

но оно жестко не проецируется. Так, в следующем примере, строки 6–7 представляют собой незаконченное перечисление мест празднования, главные фразовые акценты в этих строках реализуются как ИК-6 с растянутой ударной гласной. В принципе, ни лексико-грамматическое, ни просодическое оформление этих строк не проецируют их потенциальную связь со следующим фрагментом, однако и не противоречат такой связи. В той системе нотации, которая используется в корпусах [Spokencorpora 2013], такого рода омонимия дискурсивного статуса делает допустимыми два варианта нотации в позиции перед А, ср. (20а) — знак «…» плюс заглавная буква, т. е. завершенность перед А, иллокутивная граница:

(20а) FS_11-m_Sp

6. ··(0.34) Отмечали мы сначала в /–одно-ом месте,,,
7. потом в /–друго-ом…
8. А потом в итоге закончилось отмечанием в /парке на \лавочке.

и (20б) — знак «,,,» (ослабленное, внутрииллокутивное многоточие) плюс строчная буква, т. е. незавершенность перед А в составе единой иллокуции:

(20б) FS_11-m_Sp

6. ··(0.34) Отмечали мы сначала в /–одно-ом месте,,,
7. потом в /–друго-ом,,,
8. а потом в итоге закончилось отмечанием в /парке на \лавочке.

Можно заключить, что в такого рода контекстах противопоставление по просодической завершенности оказывается нейтрализованным.

Другой класс проблемных случаев связан со сменой типа иллокуции на границе перед А. Сочетаемость иллокутивных значений вопроса, директива, обращения, других частных типов иллокуций с дискурсивной незавершенностью пока плохо изучена, поэтому не всегда удается однозначно квалифицировать наблюдаемые просодические паттерны как завершенные или как незавершенные. Приведем пример, где во фрагменте, предшествующем А, на иллокуцию сообщения наложена просодически ярко выраженная эмфаза:

(21) FS_18-f_Sp

269. ···· Брентон так на него /смотрит,
270. ·· и \говорит:
271. ··· «Mein \Mutter!»
272. ··· И \проходит!
273. ···· А водитель смотрит на \–меня,

Здесь восклицание, предшествующее А, реализуется нестандартно — с резким падением из высокого регистра на предударном слоге, с последующим пологим падением на ударном и слабым пологим подъемом на заударном, см. **рисунок 1**. Можно усмотреть здесь наложение значения незавершенности,

однако объективировать такое решение крайне трудно. Поэтому данный случай и ему подобные мы отнесли с спорным:

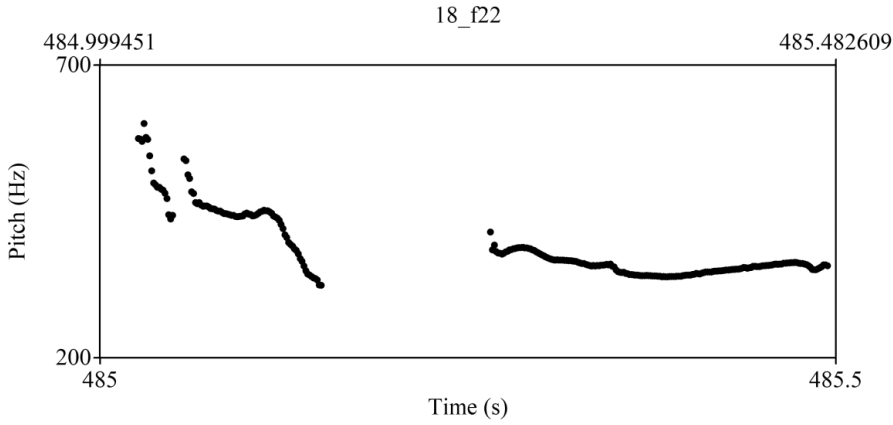


Рисунок 1

Наконец, очевидным образом, группу проблемных случаев пополняют контексты, где в просодическую реализацию А-конструкций вмешивается речевой сбой:

(22) FS_14-f_Sp

50. ну в общем ·· мы стояли-/стоя̲ли,
51. и ==
52. ·· а я-то всё /понима̲ю,

Таковы, в самом сжатом изложении, ситуации, в которых фрагмент, предшествующий А, может быть охарактеризован как просодически завершенный, как просодически незавершенный или однозначная характеризация его просодического статуса наталкивается на определенные трудности. Обратимся теперь к некоторым общим количественным наблюдениям.

3. Конструкции с *А* в устном подкорпусе в сопоставлении с письменным: некоторые количественные данные и выводы

3.1. Конструкции с *А* в устных и письменных версиях рассказов

Для того чтобы сравнить стратегии использования А-конструкций, письменные и устные версии рассказов были выровнены, с тем, чтобы выявить пары текстовых фрагментов (устный/письменный), отсылающих к одному и тому же событию в сюжетной канве рассказа. Для двадцати одной конструкции (чуть меньше

трети из 76 обнаруженных в письменном подкорпусе) нашлись прямые аналоги в письменной версии. В письменных версиях рассказов решения о знаках пунктуации принимали сами испытуемые, поэтому есть основания рассматривать «запятую плюс строчное *а*» как аналог просодической незавершенности, а «точку (вопросительный знак, двоеточие, кавычки) плюс заглавное А» — как аналог просодической завершенности. Выяснилось, что во всех обнаруженных парах, за исключением одной, стратегия выбора просодии\знака перед А в устной и письменной версиях совпадала. В (23а,б) и (24а,б) приведены два из двадцати случаев совпадения стратегии, в (25а,б) — единственный пример несовпадения (устная версия, как и во всех прочих примерах, приводится в корпусной транскрипции, а в письменной версии сохранена орфография и пунктуация испытуемого):

(23а) FS_11-m_Sp

39. /поэтому ·· можешь пока потусоваться /дома-а,
40. ·· а /потом /наутро /–приедешь в больницу^{w},,,
41. ··мы тебя сразу ·· /пропишем,
42. и \прооперируем.»

(23б) FS_11-m_Wr

Так что я могу пока поехать домой, а на утро вернуться в больницу, где меня сразу зарегистрируют и прооперируют.

(24а) FS_05-m_Sp

45. ··· и /тут папа \думает:
46. «А’ || а’ || а собственно ·· ээ так ли велика разница между шестым и \шестнадцатым?»

(24б) FS_05-m__Wr

И папа подумал: «А почему, собственно, 16-е? Здесь же ясно написано: шестое, разве нет?».

(25а) FS_11-m_Sp

103. ···\думаю:
104. «/Ну \всё,
105. не могу я больше /лежать,
106. нужно чего-то \делать.»
107. А уже время одиннадцать /часов,
108. ·· ээ все легли /спат-ть,
109. тихий /час,
110. я \думаю:
111. «\О’кей,
112. у меня есть /книжечка,
113. пойду почитаю в \холле.»

(25б) FS_11-m_Wr

К вечеру лежать мне надоело, а так как было уже после одиннадцати, и все спали, мне пришлось выйти читать в холл.

Возможное объяснение наблюдаемого параллелизма состоит в том, что говорящий, выбирая определенный способ упаковки фрагментов ситуации и тип связи между этими фрагментами, выбирает и наиболее приемлемый — с его точки зрения — для данного контекста уровень интеграции фрагментов в единое целое. Совпадение контекстов приводит к совпадению уровня интеграции.

3.2. *А* и *НО*

Напомню, что в устном подкорпусе было задокументировано 93 вхождения А-конструкций, в письменном — 76 вхождений. Эти данные демонстрируют следующую важную тенденцию: частотность А-конструкций не зависит от регистра речи, а определяется преимущественно сюжетом нарратива: в письменном подкорпусе обнаруживается доля А-конструкций, сопоставимая с устным, но чуть более высокая — 93 на 10 000 слов (9,3 на 1000 слов) в устном против 76 на 7000 слов (10,8 на 1000 слов) в письменном. Сравним эти цифры с данными по тому же корпусу «Веселых историй из жизни», приведенными в Подлесская 2016 для союза НО, см. **Таблицу 1**:

Таблица 1 (на основе Подлесская 2016)

Корпус	Число слов в корпусе	НО всего	НО/ 1000 сл.	А всего	А / 1000 сл.
Веселые истории из жизни, устные	10 000	53	5,3	93	9,3
Веселые истории из жизни, письменные	7 000	31	4,4	76	10,8

Очевидно, что НО в обоих регистрах существенно более редкий маркер, чем А, и его частоты, также, как и частоты А, мало разнятся в устных и письменных версиях (НО — незначительно чаще в устных рассказах, А — незначительно чаще в письменных). Возможное объяснение большей частотности А кроется в том, что в значении НО более выражен пропозициональный компонент (условно «несоответствие ожиданиям»), тогда как значение А — гораздо более общее и связано, прежде всего, с организацией дискурса.

По-видимому, это же различие в значении приводит и к тому, что А и НО демонстрируют разные распределения по завершенности/незавершенности первого компонента конструкции. Сравним сначала данные **Таблицы 2** и **Таблицы 3**. Здесь мы видим, что в устном подкорпусе доля просодически завершенных компонентов перед А, несколько выше, чем перед НО, но это разница незначительна, и, в целом, доли в колонках «завершенность», «незавершенность» и «спорные случаи» условно укладываются в пропорцию 40/40/20:

Таблица 2 (на основе Подлесская 2016)

Корпус	НО, всего	просодическая завершенность перед НО (доля, %)	просодическая НЕзавершенность перед НО (доля, %)	Спорная просодия (доля, %)
Веселые истории из жизни, устные	53	21 **(39,6)**	22 **(41,5)**	10 **(18,9)**

Таблица 3

Корпус	А, всего	просодическая завершенность перед А (доля, %)	просодическая НЕзавершенность перед А (доля, %)	Спорная просодия (доля, %)
Веселые истории из жизни, устные	93	41 **44,1%**	35 **37,6%**	17 **18,3%**

В письменном подкорпусе при сравнении пунктуационных аналогов завершенности\незавершенности первого компонента картина иная, ср. Таблицу 4 и Таблицу 5. В НО-конструкциях доля незавершенных первых компонентов почти в три раза превосходит долю завершенных, тогда как для А-конструкций доля незавершенных лишь незначительно превосходит долю завершенных. Соответственно, доля завершенных для А-конструкций почти в два раза выше доли завершенных для НО-конструкций:

Таблица 4 (на основе Подлесская 2016)

Корпус	НО, всего	НО с заглавной буквы (доля, %)	НО не с заглавной буквы (доля, %)
Веселые истории из жизни, письменные	31	8 **25,8%**	23 **74,2%**

Таблица 5

Корпус	А, всего	А с заглавной буквы (доля, %)	А не с заглавной буквы (доля, %)
Веселые истории из жизни, письменные	76	34 **(44,7%)**	42 **(55,3%)**

По-видимому, для НО-конструкций в письменных текстах более влиятельным оказывается прототип с двухместным союзом НО, где оба компонента объединены в единую иллокуцию, и запятая является стандартным пунктуационным проектором, «обещающим» читающему появление в тексте фрагмента, связанного с текущим. Имеющиеся в устных текстах конструкции с просодически завершенным компонентом перед НО строятся по модели парцелляции второго компонента, ср. строки 7–8 и 9–10 в (26):

(26) FS_02-f_Sp

7. И-и /решили \/–мы ··(0.39) мм(0.28) /поехать на смотровую \ площ<u>а</u>дку^{w}.
8. ··(0.31) {ЧМОКАНЬЕ 0.14} Но там \н<u>е</u> было парковочных мест.
9. /Мы решили заехать на /территорию \университ<u>е</u>та^{w}.
10. ··(0.31) Но /там висит ↑\–**кирп<u>и-и</u>ч**.

В письменных текстах авторы стараются избегать таких построений, ориентируясь на стереотипы прескриптивной пунктуации. В отличие от НО, для дискурсивного маркера А в письменных текстах сохраняется два формата — наряду с двухместным А сохраняется хождение одноместного А, ср. (24б). Вполне вероятно, что двухместные употребления А с парцелляцией подвергаются такой же пунктуационной коррекции, как и двухместные НО, однако одноместные А с завершенным первым компонентом сохраняются. Поэтому общая доля А-конструкций в завершенным первым компонентом в письменных текстах сохраняется столь же высокой, что и в устных текстах.

Разумеется, обследованный корпус имеет небольшой объем для полновесных количественных выводов, однако он дает эмпирическую основу для качественного анализа, позволяющего сложить общий портрет дискурсивного маркера из разнообразия его лексико-грамматических, семантических и просодических проявлений. Дальнейшее расширение просодически размеченного массива данных позволит усилить предложенную функциональную аргументацию и повысить статистическую валидность материала.

Литература

1. *Брызгунова Е. А.* (1982) Интонация, Русская грамматика, том 1, М.: Наука, 103–118.
2. *Грамматика-80* (1982): Русская грамматика. В 2-х тт. / Н. Ю. Шведова (гл. ред.). М.: Наука
3. *Зализняк Анна А., Микаэлян И.* (2005) Русский союз А как лингвоспецифическое слово. Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции «Диалог' 2005» по компьютерной лингвистике и ее приложениям. Звенигород.
4. *Зализняк Анна А., Микаэлян И.* (2018) Русское А: новый взгляд на старую проблему. Russian Linguistics. (in print)

5. *Йокояма О.* (1990) К анализу русских сочинительных союзов. . В.: Арутюнова Н. Д. (ред.) Логический анализ языка: противоречивость и аномальность текста.М.: Наука, 190–193.
6. *Кибрик А. А., Подлесская В. И.* (ред.) (2009) Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М.: ЯСК.
7. *Крейдлин Г. Е., Падучева Е. В.* (1974а) Значение и синтаксические свойства союза А. Научно-техническая информация, 2–9, 31–37.
8. *Крейдлин Г. Е., Падучева Е. В.* (1974б) Взаимодействие ассоциативных значений и актуального членения в предложениях с союзом А. Научно-техническая информация, 2–10, 32–37.
9. *Левин Ю. И.* (1970) Об одной группе союзов русского языка. Машинный перевод и прикладная лингвистика, 13, 64–88.
10. *Падучева Е. В.* (1996а) Субъективная модальность: иллокутивные показатели и вводные слова // Падучева Е. В. Семантические исследования. М.: Школа «Языки русской культуры», 297–320.
11. *Подлесская В. И.* (2016) «Но по расчету по моему должна родить»: конструкции с союзом но по данным корпусов с просодической разметкой. Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог» (2016). Выпуск 15, 561–565
12. *Санников В. З.* (2008) Русский синтаксис в семантико-прагматическом пространстве. М.: ЯСК.
13. *Урысон Е. В.* (2006) Подсистема русских сочинительных союзов И, А и НО. Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции «Диалог' 2006» по компьютерной лингвистике и ее приложениям. Бекасово, 519–526.
14. *Урысон Е. В.* (2011) Опыт описания семантики союзов. М.: ЯСК.
15. *Янко Т. Е.* (1990) Еще раз о союзах А и НО. В.: Арутюнова Н. Д. (ред.) Логический анализ языка: противоречивость и аномальность текста. М.: Наука, 246–258.
16. *Янко Т. Е.* (2008) Интонационные стратегии русской речи в сопоставительном аспекте. Москва: Языки славянских культур.

References

1. *Bryzgunova E. A.* (1982) Intonation [Intonatsiya], Russian Grammar [Russkaja grammatika]. Vol. 1, Nauka, Moscow, pp. 98–118
2. *Janko T. E.* (2008) Intonacionnye strategii russkoj rechi v tipologicheskom aspekte [Intonational strategies in spoken Russian from a comparative perspective]. Moskva: Jazyki Slavjanskix Kul'tur.
3. *Kibrik A. A., Podlesskaya V. I.* [Eds.] (2009) Rasskazy o snovidenijax: korpusnoe issledovanie usntogo russkogo diskursa [Night Dream Stories: A corpus study of spoken Russian discourse]. Moscow: Jazyki Slavjanskix Kul'tur.
4. *Lobke Aelbrecht, Haegeman Liliane, Rachel Nye* (Eds.) (2012) Main Clause Phenomena: New Horizons [Linguistik Aktuell/Linguistics Today, 190] John Benjamins

5. *Malchukov A.* (2004) Towards a Semantic Typology of Adversative and Contrast Marking. Journal of Semantics 21, 177–198
6. *Podlesskaya V.* (2016) "No po raschotu po moemu dolzhna rodit": the russian conjunction NO viewed through the prism of prosodically annotated corpus data. Computational Linguistics and Intellectual Technologies. Proceedings of the Annual International Conference "Dialogue" (2016) Issue 15, 561–565
7. Spokencorpora (2013) Prosodically Annotated Corpus of Spoken Russian (PrACS-Russ). Pilot version. Online: http://spokencorpora.ru

REFERRING EXPRESSION GENERATION FOR QUESTION ANSWERING AND GRAPH VISUALIZATION[1] [2]

Rygaev I. P. (irygaev@gmail.com)

Laboratory of Computational Linguistics, A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia

This paper describes a practical solution for the task of referring expressions generation (REG) in the context of a question-answering system. When an answer to a question is found in the knowledge base the system has to decide how to present the answer to the user, which properties uniquely distinguish the object found from other objects in the knowledge base. Another task where referring expressions would be useful is the semantic graph visualization task. Building on top of the graph-based approach presented by Krahmer et al in 2003 this paper provides some practical improvements to the algorithm, namely: 1) Instead of depth-first graph search we use breadth-first search, which is dramatically faster when a scene graph is big but the description graph to be found is small, 2) Limit on the size (the number of edges) of the resulting description graph to increase performance and avoid useless long descriptions. Also a sketch on linguistic realization of the referring expressions is outlined.

Keywords: natural language generation, referring expression generation, question answering, semantic graph, semantic web, inference

ГЕНЕРАЦИЯ РЕФЕРЕНЦИАЛЬНЫХ ВЫРАЖЕНИЙ ДЛЯ ОТВЕТОВ НА ВОПРОСЫ И ВИЗУАЛИЗАЦИИ ГРАФОВ

Рыгаев И. П. (irygaev@gmail.com)

Лаборатория компьютерной лингвистики Института проблем передачи информации им. А. А. Харкевича РАН, Москва, Россия

[1] This paper presents the results of a joint effort of the team of the Laboratory of Computational Linguistics of the Institute for Information Transmission Problems of the Russian Academy of Sciences. The team includes I. Boguslavsky, L. Iomdin, A. Lazursky, S. Timoshenko, T. Frolova, V. Dikonov, E. Inshakova, V. Sizov and others. I would like to thank my colleagues for their wonderful collaboration.

[2] This work was supported by the RSF grant 16-18-10422, which is gratefully acknowledged.

1. Introduction

The semantic text analyzer SemETAP, under development in the Laboratory of Computational Linguistics of IITP RAS, is aiming at modelling deep understanding of natural language texts (in Russian). The analyzer includes a powerful linguistic processor and various linguistic and extra-linguistic resources—a combinatorial dictionary, an ontology, a repository of individuals, a set of inference rules and an inference engine [Boguslavsky 2011]; [Boguslavsky et al 2010, 2013]. One of the applications of SemETAP is a question-answering system able to answer questions for which there is no direct answer in the original text [Boguslavsky et al 2015]; [Rygaev 2017]. For example, given the sentence:

(1) Зенит не смог спасти матч
Zenit could not save the match

The system can answer:

(2) Кто проиграл?
Who has lost the match?

In order to get the answer the following steps are performed:

1. A language-independent basic semantic structure (BSemS) of the first sentence is built. BSemS consists of a set of binary predicates (RDF triples) and can be seen as a semantic graph where nodes correspond to individuals mentioned in the sentence (including event individuals) and arcs correspond to relations between the individuals.
2. Inference rules are applied to extend BSemS adding new individuals and relations and thus forming an enhanced semantic structure (EnSemS). In our example this step (among other things) adds the knowledge that Zenit has lost the match.
3. BSemS of the question is build. It is similar to that of the affirmative sentence but wh-words are marked in a special way.
4. The question BSemS is used as a pattern to search within the semantic graph of the text. The search returns individuals (graph nodes) corresponding to the wh-words in the question.
5. Along with the node ID certain meaningful information is returned such as the type of the found individual and its name (if exists).
6. Based on this information a linguistic representation of the answer is build and inserted into the text of the question instead of the wh-word, thus generating the answer sentence. Then the answer sentence undergoes some slight modifications (such as agreement and word order change) and is presented to the user. In our example the resulting sentence would be:

(3) Проиграл футбольный клуб «Зенит»
Football club Zenit has lost the match

As mentioned in p. 5 only a few relations (mainly type and name) are currently used to generate the referring expression for the answer. In case the text does not contain the name of the team we are left only with its type (football club) which is not distinguishing enough as can be seen in (4):

(4) Аршавин не смог спасти матч. Кто проиграл?
Arshavin could not save the match. Who has lost the match?

In this case we would like to have the answer *Ashavin's team* or *Arshavin's football club*[3]. The answer *The football club* will not be distinguishing as there are two football clubs in the match. Moreover the type of the potential candidates for the answer is already presupposed by the question (assuming we are talking about a football match). So such an answer does not provide any new information.

The goal of this paper is to present a solution for the general content selection task for referring expression generation (REG). The algorithm should find a minimal distinguishing description of a node in a graph taking into account all existing relations. The second stage (linguistic realization) is beyond the scope of this paper though a sketch of how the problem can be attacked will be outlined.

The paper is organized as follows: Section 2 poses a problem of referring expression generation for question answering, Section 3 presents the graph visualization problem as another task which would benefit from the REG solution, Section 4 discusses related work and existing algorithms for REG including a graph-based approach, Section 5 describes our practical improvements to the graph-based algorithm, Section 6 discusses the evaluation of the new algorithm, Section 7 outlines a sketch of how linguistic realization of the referring expression can be generated, and Section 8 concludes the paper.

2. Referring expressions for answers

Referential choice is known to be a multi-factor probabilistic process [Kibrik et al 2010]. An individual can be referred in discourse by a pronoun, a proper name or a common name potentially modified by an adjective, prepositional phrase or relative clause. Reference also can be of different types such as specific/generic, singular/plural, definite/indefinite and so on.

In what follows we limit ourselves only to specific singular reference. This follows from the nature of the question-answering system. An answer to a question that the system is able to produce is always a specific individual from the knowledge base. In case there are many answers the system will just list them all one by one. Also the usage of pronouns is not an option because the system currently does not take into account the preceding discourse (each question-answer pair is considered to be a separate conversation). Hence ultimately a noun phrase must be generated. But the aim of this paper is limited only to selection of the content for the noun phrase generation.

The problem can be stated as follows:

(5) Given a target node in a semantic graph (called scene graph) find a minimal subgraph (called description graph) which uniquely distinguishes the target node from any other nodes in the graph.

[3] Of course we can look up in the repository of individuals, find out the name of Arshavin's team and use it for the answer. But let's assume the repository of individuals is not available or the team is not there. Anyway, the task of extending the knowledge graph from RI is independent of the task of referring expression generation which is the aim of this paper.

We assume that the information from the description graph will be enough to generate linguistic realization of the referring expression.

Statement (5) covers all our limitations and is generic enough to capture also referring expressions in graph visualization (see the next section). But for question answering certain additional considerations should be taken into account. The content selected for the answer should not include the information which was used to find the node itself. In other word it should not include the presuppositions of the question. Consider the following question-answer pair:

(6) Who won? The winner

This answer is obvious and useless since it is already presupposed by the question itself. To avoid such answers we need to exclude question presuppositions from the scope of search for a description graph. By doing so we need to take into account not only the basic semantic structure of the question but also all the inferences which can be made out of it. Consider the next example:

(7) Who bought the car?
a. The buyer
b. The one who paid
c. The one who received the car

All three answers are useless though only (7a) is contained within the basic semantic representation of the question, and (7b-c) are not contained but rather entailed by it.

So when generating referring expressions for answers we need first to produce an EnSemS of the question (applying inferences to BSemS) and remove the resulting EnSemS from the scene graph before searching for a description graph. In addition to solving the problem with useless answers this scene graph reduction will help with the performance of the REG algorithm.

3. Referring expressions in graph visualization

Another task where short unique referring expressions would be useful is the visualization of a complex semantic graph. In the graphs which are built by SemETAP each node has a unique ID which contains a type of the individual and a certain number postfix. When there are multiple nodes of the same type it is hard to follow which particular individual a node represents.

This is especially problematic for auxiliary nodes such as **`Complete`** or **`EpistModality`**. **`Complete`** nodes are usually generated from perfective aspect of a verb and represent the completion of an event. **`EpistModality`** nodes represent the degree of confidence in a proposition and are attached to any event which is stated or inferred to be a true fact. There can be a lot of auxiliary nodes in the graph and in order to distinguish between them a user needs to check adjacent nodes, sometimes several spans in different directions, which is cumbersome and time-consuming. A unique descriptive expression as a node ID would be really helpful.

See below a graph for a sentence:

(8) Иван купил зонтик у Петра
Ivan bought an umbrella from Peter

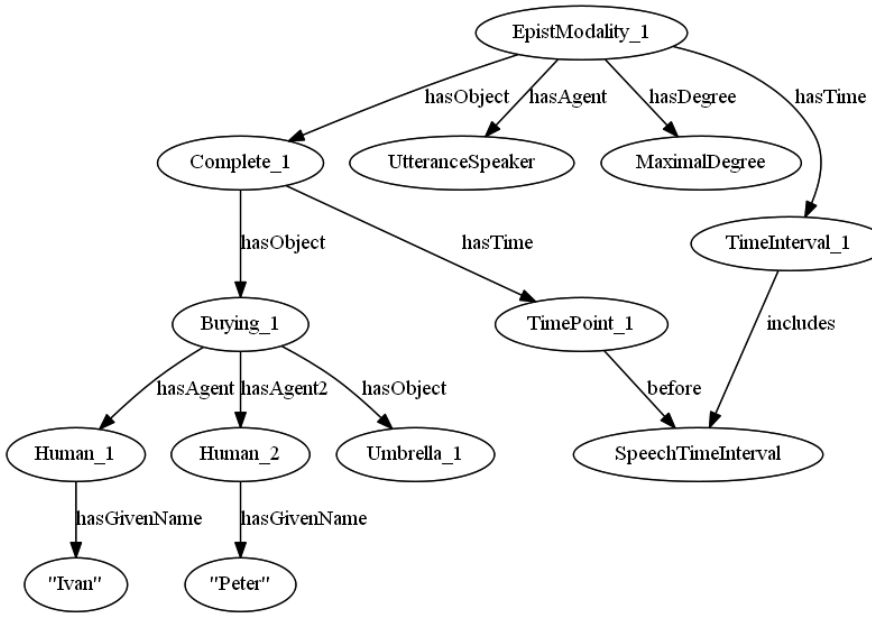


Fig. 1. Basic semantic graph for a sentence 'Ivan bought an umbrella from Peter'

This graph is fairly clear since it is rather small. But when we add inferences to it, the graph becomes unreadable. Instead of providing a picture of it we will list statistics of its node types.

Table 1. Node type statistics of the enhanced semantic graph for the sentence 'Ivan bought an umbrella from Peter'

Grouping	Node type	Number of nodes	Explanation
Objects (4)	Human	2	Two persons—Ivan and Peter
	Umbrella	1	An umbrella
	Currency Measure	1	Money
Events (13)	Buying	1	Ivan bought the umbrella from Peter
	Selling	1	Peter sold the umbrella to Ivan
	Payment	1	Ivan paid for the umbrella to Peter
	Exchange	2	Ivan exchanged the money for the umbrella Peter exchanged the umbrella for the money
	Giving	2	Ivan gave the money to Peter Peter gave the umbrella to Ivan
	Getting	2	Ivan got the umbrella from Peter Peter got the money from Ivan
	Own	4	Ivan owned the money before the purchase Peter owned the umbrella before the purchase Ivan owns the umbrella after the purchase Peter owns the money after the purchase
Auxiliary (42)	Complete	9	Completion for each event except ownership
	EpistModality	22	Facticity of each event and each completion
	TimeInterval	8	Time positions of the events. For some events
	TimePoint	3	they coincide.

For certain nodes (**`Umbrella`**, **`Buying`**, etc.) their class is distinguishing enough, other nodes (people) can be identified by proper names. But when we come to non-unique event types, some descriptive content (such as event arguments) is required to distinguish between them[4]. And for auxiliary nodes we need to include arguments of their arguments as well.

Similar problems arise when one tries to browse open knowledge bases in Semantic Web. If DBpedia (dbpedia.org, [Auer et al 2007]) uses descriptive URIs inherited from Wikipedia article names, Wikidata (www.wikidata.org, [Vrandečić and Krötzsch 2014]) abandons this notation for the sake of multilingualism and uses numeric object IDs such as Q175117. Especially cumbersome is the data structure in BabelNet (babelnet.org, [Navigli and Ponzetto 2012]). The picture below shows how a semantic concept of Apple (fruit) is presented in their linked data interface:

[4] As one may notice node naming is not the only problem in the complex graph visualization. Other issues include proper arrangement of the nodes and the ability for a user to interact with the graph. But even if those two issues are resolved poor node naming will prevent the user from reading and understanding the graph quickly. So we will concentrate on the node naming as this is the only linguistic task in graph visualization.

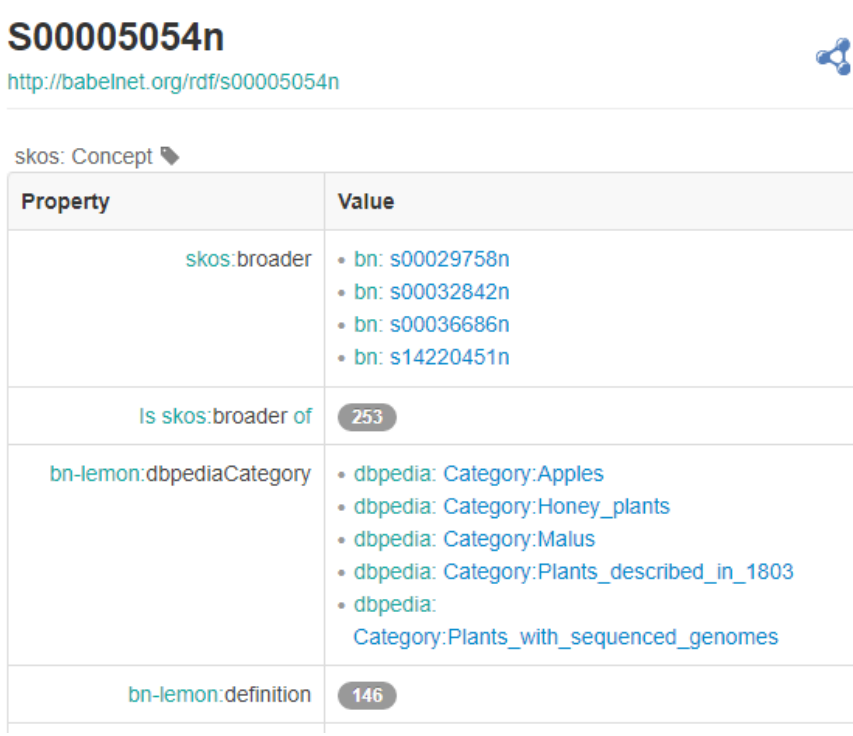


Fig. 2. Semantic concept of Apple (fruit) in the BabalNet linked data interface

If one wants to understand what the broader concepts are, they have to navigate to a particular concept, look at the definition attribute which is also not descriptive but refers to an object named something like s00005054n_Gloss1_EN. And only after navigating to this BabelGloss object one can find a definition of the concept. Adding automatically generated meaningful descriptions instead of numeric concept IDs (or in addition to them) would make the property sheet much clearer.

SPARQL[5] specification [Prud'hommeaux, Seaborne 2008] contains a DESCRIBE query which should return an RDF graph which describes a particular resource (or resources) in an RDF storage, but it is left up to the server to decide which triples to include into the description. This would be another good application for referring expression generation in the context similar to the graph visualization.

4. Existing algorithms for REG

The history of research on the referring expression generation [cf. Krahmer and van Deemter 2012] goes back to [Winograd 1972], who first presented a primitive algorithm for naming objects and events. Since then a number of algorithms have been suggested and evaluated. We briefly discuss the major ones:

[5] A query language for RDF knowledge bases in Semantic Web.

1. *Full Brevity* algorithm [Dale 1989] guarantees to generate the shortest possible distinguishing description. First it tries every single property of the target and checks if it alone rules out all the distractors. If that fails it then tries all possible combinations of *two* properties, then *three* properties and so on until a distinguishing description is found or all the properties are exhausted. This algorithm is computationally expensive and surprisingly of low human-likeness. It was shown that human speakers often produce non-minimal descriptions [Pechmann 1989]; [Engelhardt et al 2006].
2. *Greedy Heuristics* algorithm [Dale 1989, 1992] is more efficient then Full Brevity. It incrementally adds one property to the description—the one which rules out most of the current distractors. Because of its incremental nature (once the property is added it is never removed) it does not always produce the shortest descriptions.
3. *The Incremental Algorithm* [Reiter and Dale 1992] is probably the most influential algorithm in REG. It is similar to Greedy Heuristics but instead of selecting properties based on their discriminating power it uses predefined preference order of the properties. It was shown that speakers prefer certain properties over others when referring to objects [Pechmann 1989]. This algorithm is of polynomial complexity and produces the most natural human-like descriptions. But it requires a preference order to be carefully specified upfront.

It needs to be pointed out that these algorithms originally were tested in a simplified set-up where objects are characterized by their properties only, but not by relations between them [Krahmer and van Deemter 2012: 181]. In our model where almost all properties are in fact relational (even object type is a relation between an individual and a class) this limitation needs to be lifted[6].

There were a number of attempts to adapt the Incremental Algorithm for relational properties [Horacek 1996]; [Krahmer and Theune 2002]; [Kelleher and Kruijff 2006] but unlike simple properties relations do not fit very well into an incremental paradigm. No one would produce a description *'the dog next to the tree in front of the garage'* when *'the dog in front of the garage'* would suffice [Krahmer et al 2003: 57].

[Krahmer et al 2003] suggests a graph-based approach for REG which covers the case of relational properties and fits very well in our knowledge representation framework (since it is already graph-based). They present a branch and bound algorithm [Land and Doig 1960] for finding a relevant subgraph with a cost function to guide the search. Roughly at each step this algorithm enumerates the neighbor edges of the current candidate description graph, checks whether adding an edge will result in a subgraph cost not exceeding the cost of the current best subgraph (if it exists) and if it so checks whether the new candidate rules out all the distractors. If the check is successful then the current best subgraph is updated and the algorithm backtracks, otherwise the same steps are performed on the new candidate.

---

[6] Other limitations such as singular references only, crisp and not vague properties only, ignoring salience in context, etc. [Krahmer and van Deemter 2012: 181] still apply to our work as well. Lifting them is the topic of future research.

Authors argue that their approach (with certain modifications) can mimic the results of all the three algorithms described above. If the cost of adding each edge and each node is the same the algorithm will produce a minimal description as Full Brevity does. If the enumeration of the neighbor edges are performed in a certain order (either based on discriminating power or predefined preference) and the first found description is returned then the results of Greedy Heuristics or the Incremental Algorithm are obtained.

We tried to apply the graph-based algorithm to our tasks. The next section describes some improvements that we had to introduce to it to make the algorithm more practical.

5. Our method

The graph-based algorithm as presented in Krahmer et al 2003:62 is recursive, i.e. it realizes a depth-first search. Starting from one relation the algorithm first explores the whole branch associated with it as far as possible before even trying another single relation. This is not an optimal strategy since we expect a useful referring expression to be relatively short.

Searching for long description (before trying all the shorter ones) can dramatically increase the time required to find a solution if the algorithm happens to take at first the wrong path. Firstly the longer the description (up to a certain threshold) the more its potential for branching, the more new candidates it produces on the next step. And secondly finding distractors for longer descriptions is also more time-consuming. Much more descriptions multiplied by much more time for checking each description makes the algorithm impractical. In our tests the depth-first algorithm could go as far as several dozen relations (still not finding a solution) when a unique description to be found was only several relations long.

To solve this problem we propose a *breadth*-first modification of the graph-based algorithm which (like Full Brevity) first tries every single relation as a full description then every combination of two relations and so on. If there is a relatively short description to be found then the breadth-first search usually finds it much faster than the depth-first search.

But if there is no unique description available then the breadth-first algorithm is slower in arriving at this conclusion. However, to confirm that there is no unique description an algorithm anyway would need to test the longest possible description graph. If the scene graph is connected (and it is usually the case) then the longest possible subgraph would be the whole scene graph. As we mentioned above exploring up to this point is not a practically available option. Also taking into account the following:

1. Long descriptions are not only time-consuming but also not very useful for a user to identify the object.
2. If a unique description does not exist the algorithm still has to return something at least partially useful. It cannot just fail or return an empty string.

We decided to introduce the length limit (in a number of edges) for a description graph. Potential descriptions of the length above the limit are not considered at all.

And when all candidates of maximal length are explored and rejected then the algorithm returns a simple subgraph containing just one edge—a type of the target node. Thus we get acceptable performance for the cost of not always finding unique descriptions (when they are sufficiently long).

In addition to that we realized that forced addition of node types to the description graph is not only beneficial for a user (it produces much more natural descriptions) but also makes the algorithm faster. This is probably due to the fact that our graphs usually contain many edges with the same relation. So a node type plus a relation is much more distinguishing then just a relation. Hence we introduced a rule: whenever a node is added to the description graph its type edge is added automatically as well.

Also we found useful to define a cost function and enumerate neighbors based on the predefined preference order of relations (similar to the Incremental Algorithm). On the top of the preference list we have proper name relations (**`hasName`**, **`hasGivenName`**, etc.) followed by argument relations (**`hasObject`**, **`hasAgent`**, etc.) and so on. This also increases the performance of the algorithm.

6. Evaluation

There are two types of evaluation that can be performed against a REG algorithm—human evaluation of the generated expressions and performance evaluation.

Human evaluation concerns how natural a referring expression is and how helpful it is to identify the target object. Without linguistic realization this type of evaluation cannot be fully performed. But some preliminary validation tests can be made. While surface realization is in progress the resulting expression is presented in a formal language called Etalog which is the language of SemETAP inference rules. It was designed in such a way as to be understandable for linguists without special mathematical or computer-science training. The full description of Etalog is out of scope of this paper. In Fig. 3 we present some examples of Etalog referring expressions in the tree view graph visualization for sentence (4). We hope that they are rather clear and self-explaining.

For evaluation we selected 51 sentences from the corpus of the football high spots (those which were not previously used to developed and test the generation of referring expressions), manually created meaningful questions to these sentences and presented the Etalog answers to four linguists familiar with Etalog asking them to evaluate informativeness and naturalness (human-likeness) of the generated referring expressions. A total of 287 question-answer pairs were evaluated.

Both characteristics were evaluated using a binary scale (yes/no). The informants were instructed to regard an answer as informative if they can unambiguously identify the referred individual within the context of the sentence based on the Etalog expression provided, and the referring expression contains new information (other than what is presupposed by the question itself). And they were instructed to regard the answer as natural if they can imagine someone using such an expression to answer this particular question within the context of this particular sentence.

There were 7 types of referred individuals in the answers. The statistics for each type (as well as the totals) is presented in Table 2. Statistics for persons (football players) is split into two parts: those identified by name and the rest.

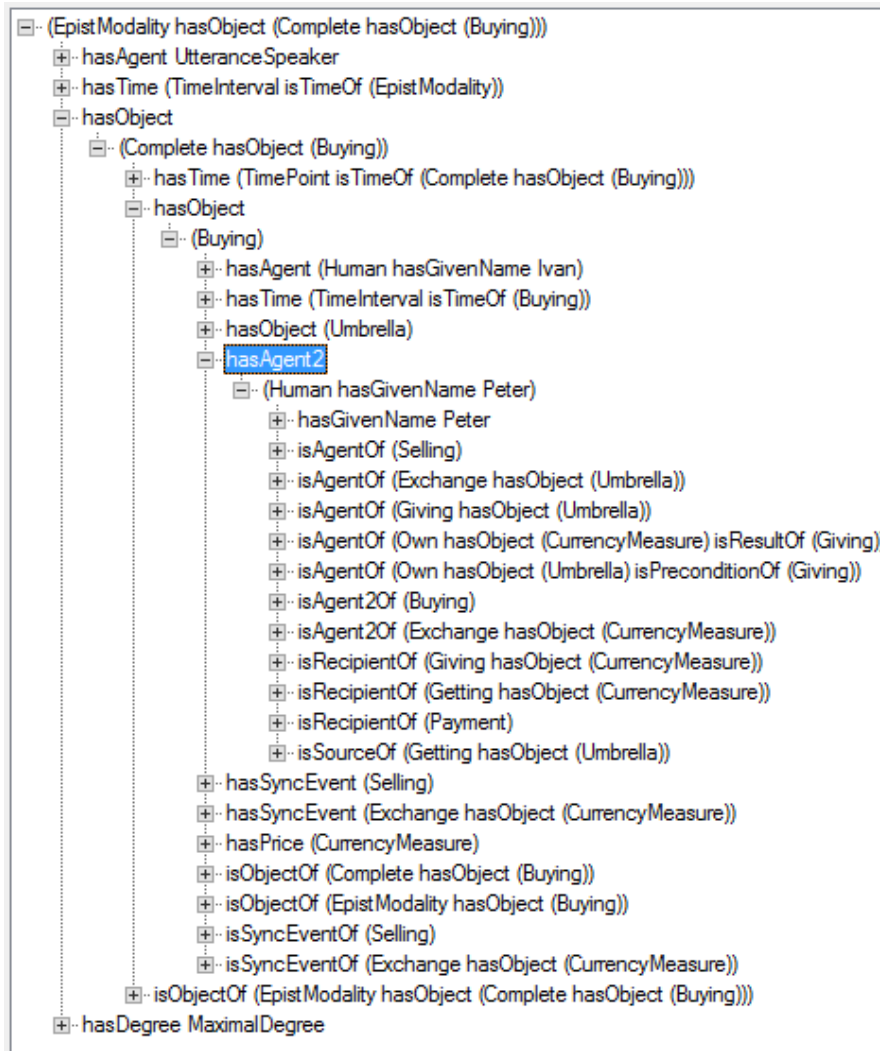


Fig. 3. Tree view of the enhanced semantic graph for the sentence 'Ivan bought an umbrella from Peter'

Table 2. Informativeness and naturalness of the answers by the referred individual type

Referred individual type	Number of answers	Informativeness	Naturalness
Person (identified by name)	81	100.00%	98.42%
Person (identified by other means)	28	36.04%	18.02%
Football team	75	28.16%	17.20%
Place (penalty area, goal area, etc.)	41	62.73%	41.61%
Event (football pass or shot)	32	58.59%	47.66%
Time	22	19.32%	3.41%
Ball	7	100.00%	42.86%
Body part	1	100.00%	50.00%
Total	**287**	**59.27%**	**47.04%**

Table 3 below displays the informative natural, informative unnatural and uninformative examples for each type of referred individuals (where applicable):

Table 3. Good and bad examples for all types of referred individuals

#	Context sentence	Question	Answer
Informative and natural answers:			
(9)	Все тот же Аппаев не попадает даже в створ ворот. All the same Appayev does not even hit the target.	Кто бьёт? Who shoots?	**(Human hasFamilyName "Аппаев")** Appayev
(10)	Думбия и Хонда выводят Мусу к воротам Малафеева, и нигерийцу оставалось лишь не промахнуться. Doumbia and Honda lead Musa to Malafeev's goal, and the Nigerian had only not to miss.	Кому оставалось лишь не промахнуться? Who had only not to miss?	**(Human livesIn Nigeria)** The Nigerian
(11)	После навеса в штрафную в исполнении Кержакова Аршавин блестящим ударом в падении вколачивает мяч в сетку. After a pass by Kerzhakov into the penalty area Arshavin with a brilliant shot in the fall hammers the ball into the net.	За какую команду играет Аршавин? Which team does Arshavin play for?	**(FootballTeam isObjectOf (PlaysFor hasAgent (Human hasName "Кержаков")))** The team which Kerzhakov plays for
(12)	А уже на последней минуте первого тайма Дзагоев не попал в створ ворот из выгодной позиции, пробив рядом со штангой. And in the final minute of the first half Dzagoev missed the target from a vantage point, shooting near the post.	Куда пробил Дзагоев? Where did Dzagoev shoot?	**(Region differentFrom (GoalArea))** Off the goal

#	Context sentence	Question	Answer
(13)	Подача в штрафную Шунина завершается опасным ударом головой Натхо, но голкипер на месте. The feed into Shunin's penalty area ends with a dangerous header by Natkho, but the goalkeeper is at the spot.	Каким ударом завершается подача? Which shot ends the feed?	**(FootballShot hasAgent (Human hasName "HATXO"))** Natkho's shot (The shot that Natkho made)
(14)	На исходе часа игры, Думбия, замкнув прострел Мусы, отправляет второй мяч в сетку ворот Диканя. At the end of an hour of play Doumbia, closing the pass of Musa, sends the second ball into Dikan's goal net.	Когда Думбия отправляет мяч в сетку? When does Doumbia send the ball into the net?	**(TimeInterval finishes (Hour))** At the end of an hour
(15)	Но удар Джуджака оказывается неточным, мяч проходит рядом со штангой. But Dzsudzsak's shot is inaccurate, the ball passes next to the post.	Что проходит рядом со штангой? What passes next to the post?	**(Ball)** The ball
Informative but unnatural answers:			
(16)	Муса навесил в штрафную на Хонду, тот скинул мяч Дзагоеву, который со второй попытки отправляет мяч в сетку ворот Малафеева. Musa lobbed to the penalty area for Honda, who threw the ball to Dzagoev, who at the second attempt sends the ball into Malafeev's goal net.	Кто навесил? Who lobbed (the ball)?	**(Human hasFamilyName "Хонда")** Honda (incorrect answer)
(17)	В следующей атаке хавбек исправился, замкнув в касание передачу Губочана. In the next attack the midfielder corrected himself closing in touch Gubochan's pass.	Кто исправился? Who corrected himself?	**(Human isAgentOf (Attack))** The one who attacked
(18)	Думбия вновь рвется к воротам, но вместо того, чтобы пробить самому, отдает пас на Мусу, которого опережает голкипер. Doumbia runs forth to the goal again, but instead of shooting himself, he passes the ball to Musa, which is left behind by the goalkeeper.	За какую команду играет Муса? Which team does Musa play for?	**(FootballTeam isObjectOf (PlaysFor hasSyncEvent (PlaysFor) hasAgent (Human hasFamilyName "Думбия")))** The team which Doumbia plays for at the same time when someone else plays for another team

#	Context sentence	Question	Answer
(19)	В концовке первого тайма Карсела-Гонсалес упускает очередной шанс своей команды открыть счет в матче, не попав даже в створ ворот. At the end of the first half, Carcela-Gonzalez misses his team's next chance to open the scoring in the match, not even hitting the target.	Куда не попали? What was not hit?	**(GoalArea isTerminalPointOf (GoalEvent))** The goal area where the goal is scored
(20)	Карим Бензема получил пас от Матьё Вальбуена и пробил по воротам, только вот в створ он не попал. Karim Benzema received a pass from Mathieu Valbuena and shot on goal, but he did not hit the target.	Какой сделали удар? Which shot was made?	**(FootballShot hasTerminalPoint (GoalArea isLocationOf (Arriving)))** The shot on the goal where something arrived
(21)	Валладарес переправил мяч в перекладину, от которой тот покинул пределы поля! Valladares repelled the ball into the crossbar, from which it left the field!	Что Валладарес переправил в перекладину? What did Valladares repel into the crossbar?	**(Ball isAgentOf (Leaving))** The leaving ball
(22)	Тем временем Думбия бил головой после навеса Щенникова—неточно. Meanwhile Doumbia shot with his head after Shchennikov's lob—inaccurate.	Чем бил Думбия? With what did Doumbia shoot?	**(Head isInstrumentOf (FootballShot))** The head with which the shot was made
Uninformative answers:			
(23)	Ари выполнял проникающую передачу на Эменике, тот отдал мяч дальше на ход Билялетдинову, и лишь Игнашевич успевает подстраховать голкипера. Ari performed a penetrating pass to Emenike, who gave the ball further to the course of Bilyaletdinov, and only Ignashevich manages to help the goalkeeper.	Кто получил передачу? Who received the pass?	**(Human isAgentOf (Translocation))** Someone who was moving
(24)	В течение минуты Жусилей дважды пытался пробить по воротам Беленова, но оба удара пришлись в защитника. Within a minute, Jucilei twice tried to shot on Belenov's goal, but both shots hit the defender.	По воротам какой команды пытался пробить Жусилей? On which team's goal did Jucilei try to shoot?	**(FootballTeam hasCoach (Human))** The team with a coach

#	Context sentence	Question	Answer
(25)	Ревякин спасает свою команду (сначала) после удара Кержакова, вытащив мяч из-под перекладины, (а затем и Семака). Revyakin saves his team (first) after Kerzhakov's shot, pulling the ball from under the crossbar, (and then after Semak's one).	Откуда вытаскивают мяч? Where the ball is pulled from?	**(Region isObjectOf (Below))** Below something
(26)	Полузащитник "ПСЖ" получил мяч в центре штрафной площади и вторым касанием пульнул по воротам. The midfielder of PSG received the ball in the center of the penalty area and shot on goal with the second touch.	Какой сделали пас? Which pass was made?	**(FootballPass hasLocation (Region))** The pass which is somewhere
(27)	И почти тут же Думбия имел возможность оформить "дубль", но удар у форварда явно не получился. And almost immediately Doumbia had the opportunity to make a double, but the forward's shot was obviously not good enough.	Когда Думбия имел возможность оформить "дубль"? When did Doumbia have the opportunity to make a double?	**(TimeInterval meetsTemporally (TimePoint))** The time right before some point in time

The evaluation shows that the system should be improved in a number of ways. The main problems would be the following:

1. The cost function for the algorithm needs to be configured more carefully. Often the system generated an expression which is formally distinguishing but useless from the human point of view (see examples (17), (25), (27) and others). Probably a more complex cost function is required which takes into account not only the predefined relation order but other things such as types of nodes, population of the required arguments, etc.
2. Duplicated individuals created by different rules are not always combined together by the equality (coreference) rules. This increases the number of distractors and leads to longer unnatural descriptions (see examples (18), (19), (20) and others). The logic to identify and join duplicated individuals should be improved.
3. Concept definitions do not always contain all the necessary information. For example, the definition of the football team should contain the information that it has a coach. If this was included then each football team in EnSemS would have that property and the answer in (24) would not be considered distinguishing.

It should be noted that pp. 2 and 3 above are not related directly to the referring expression generation algorithm but rather to the construction of the scene graph (EnSemS).

For performance evaluation we also present some preliminary figures. They are not final and there is still a potential for optimization. But the tendency is clear—time grows exponentially with the length of the expression. This is the reason we introduced a hard length limit to make the algorithm practically applicable.

Table 4. Average generation time and number of iterations for different description lengths

Description length	Average generation time, ms	Average number of iterations per target	Average time per iteration, ms
1	9.80	1.00	9.80
3	37.40	14.93	2.50
5	369.35	113.96	3.24
7	2,565.00	772.85	3.32

The first column in the table shows the length of the generated referring expression (in the number of edges of the description subgraph). In the majority of cases it is an odd number because edges are usually added in pairs—once a new node is added to the description its type is also added which creates an additional edge in the subgraph. Descriptions of even lengths are generated sometimes too but they do not have enough statistics, so they are omitted from the table.

The second column shows the average time (in milliseconds) required to generate an expression of the given length. The third columns displays the average number of iterations needed, i.e. the number of different descriptions tried before arriving at the solution. And the forth column shows the average time (in milliseconds) of one iteration.

It is clear from the table that the generation time growth mostly comes from the increase of the number of iterations while the average iteration time growth is rather moderate.

7. Linguistic realization

Although a full-fledged linguistic realization or referring expressions is beyond the scope of this paper we briefly present a sketch of how it could be realized.

As mentioned in the previous section we are able to generate referring expressions in Etalog formalism. An Etalog expression can serve as a template for surface realization in a natural language. Consider an example:

(28) **`(Own hasObject (CurrencyMeasure) isResultOf (Giving))`**

This can be realized in English as follows: *'The ownership of money as a result of a transfer'*. Parallels are straightforward. Roughly what needs to be done is to replace ontological concepts with corresponding words and semantic relations with syntactic ones. This process is exactly opposite to the semantic analysis which SemETAP is already capable of.

One important aspect of an Etalog expression is that it presents a description graph in a tree-like form. This tree can be used as a template for a syntactic tree of the

corresponding linguistic expression. In order to convert an arbitrary connected graph to a tree with a given head the following two steps are performed:

1. Direction of certain edges of the graph is reversed so all edges point from the head to the leaves and not vice versa. This is done through the use of inverse relations. For example, **`isResultOf`** is an inverse relation of **`hasResult`**. Whenever an unwanted incoming relation is found it can be replaced with an outgoing inverse relation.
2. Loops are eliminated. This is done by splitting a node and marking the resulting split nodes with an explicit variable. The second appearance of the variable in the expression lacks any descriptive content and can be realized as a pronoun:

(29) **`(Human ?x isAgentOf (Shaving hasObject ?x))`**
'A person who shaved (himself)'

8. Conclusion

In this paper we presented a practical realization of a graph-based algorithm for the content selection task in the referring expression generation (REG). Starting from the two needful applications of referring expressions—in the question answering and in the node naming for graph visualization, a number of practical improvements for the algorithm were suggested such as breadth-first search (instead of depth-first) and a hard limit for the description length. A preliminary evaluation for the new algorithm was provided and a sketch of the process of generating linguistic expressions based on the formal Etalog expressions was outlined.

References

1. *Auer S., Bizer C., Lehmann J., Kobilarov G., Cyganiak R., Ives Z.* (2007) DBpedia: A Nucleus for a Web of Open Data. Proceedings of ISWC 2007.
2. *Boguslavsky I. M.* (2011). Semantic Analysis Based on Linguistic and Ontological Resources. Proceedings of the 5th International Conference on the Meaning—Text Theory. Barcelona, September 8–9, 2011. Igor Boguslavsky and Leo Wanner (Eds.), p. 25–36.
3. *Boguslavsky I. M., Iomdin L. L., Sizov V. G., Timoshenko S. P.* (2010). Interfacing the Lexicon and the Ontology in a Semantic Analyzer. COLING 2010. Proceedings of the 6th Workshop on Ontologies and Lexical Resources (Ontolex 2010), Beijing, August 2010, p. 67–76.
4. *Boguslavsky I. M., Dikonov V. G., Iomdin L. L., Timoshenko S. P.* (2013). Semantic representation for NL understanding. Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2013), p. 132–144.
5. *Boguslavsky I. M., Dikonov V. G., Iomdin L. L., Lazursky A. V., Sizov V. G., Timoshenko S. P.* (2015). Semantic Analysis and Question Answering: a System Under Development. Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2015), p.62–79.

6. *Dale, R.* (1989) Cooking up referring expressions. Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL), p. 68–75.
7. *Dale, R.* (1992) Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes. The MIT Press, Cambridge, MA.
8. *Engelhardt, P. E., Bailey K. G. D., Ferreira F.* (2006) Do speakers and listeners observe the Gricean Maxim of Quantity? Journal of Memory and Language, 54:554–573.
9. *Horacek, H.* (1996) A new algorithm for generating referring expressions. Proceedings of the 12th European Conference on Artificial Intelligence (ECAI), p. 577–581, Budapest.
10. *Kelleher, J., Kruijff G.-J.* (2006) Incremental generation of spatial referring expressions in situated dialog. Proceedings of the 21st International Conference on Computational Linguistics (COLING) and 44th Annual Meeting of the Association for Computational Linguistics (ACL), p. 1041–1048, Sydney
11. *Kibrik A. A., Dobrov G. B., Zalmanov D. A., Linnik A. S., Lukashevich N. V.* (2010) Referential choice as a multi-factor probabilistic process. Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2010), p. 173–180.
12. *Krahmer, E., Theune M.* (2002) Efficient context-sensitive generation of descriptions in context. Information Sharing: Givenness and Newness in Language Processing. CSLI Publications, Stanford, CA, p. 223–264.
13. *Krahmer, E., van Deemter, K.* (2012). Computational Generation of Referring Expressions: A Survey. Computational Linguistics, 38(1), p. 173–218.
14. *Krahmer, E., van Erk, S., Verleg, A.* (2003). Graph-Based Generation of Referring Expressions. Computational Linguistics, 29(1), p. 53–72.
15. *Land A. H., Doig A. G.* (1960). An automatic method of solving discrete programming problems. Econometrica. 28 (3). p. 497–520.
16. *Navigli R., Ponzetto S.* (2012) BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, 193, Elsevier, 2012, pp. 217–250.
17. *Pechmann, Th.* (1989) Incremental speech production and referential overspecification. Linguistics, 27:98–110.
18. *Prud'hommeaux E., Seaborne A.* (2008) SPARQL Query Language for RDF. W3C Recommendation 15 January 2008. https://www.w3.org/TR/rdf-sparql-query/
19. *Reiter, E., Dale R.* (1992) A fast algorithm for the generation of referring expressions. Proceedings of the 14th International Conference on Computational Linguistics (COLING), p. 232–238, Nantes.
20. *Rygaev I. P.* (2017) Rule-based Reasoning in Semantic Text Analysis. Proceedings of the Doctoral Consortium, Challenge, Industry Track, Tutorials and Posters @ RuleML+RR 2017 hosted by hosted by International Joint Conference on Rules and Reasoning 2017 (RuleML+RR 2017).
21. *Vrandečić D., Krötzsch M.* (2014) Wikidata: a free collaborative knowledgebase. Communications of the ACM, 2014.
22. *Winograd, T.* (1972). Understanding natural language. Cognitive Psychology, 3(1).

СТРУКТУРА ПОВСЕДНЕВНОГО ДИАЛОГА КАК ПОСЛЕДОВАТЕЛЬНОСТЬ РЕЧЕВЫХ АКТОВ[1]

Шерстинова Т. Ю. (sherstinova@spbu.ru)
Филологический факультет СПбГУ;
Национальный исследовательский университет
«Высшая школа экономики»; Санкт-Петербург, Россия

Исследование структуры повседневного диалога проведено на материале 73 микродиалогов повседневной речевой коммуникации из корпуса устной русской речи «Один речевой день» (ОРД корпус). Задачей исследования было выяснение того, какие типы речевых актов чаще всего инициируют и завершают диалог, а также выявление наиболее типичных последовательностей речевых актов в структуре диалога. Была проанализирована речь 30 человек (6 информантов и 24 коммуникантов) в объеме 2230 речевых актов, относящихся как к профессиональным, так и бытовым разговорам. Для подсчета наиболее частотных последовательностей речевых актов использовалась техника n-граммного анализа. Полученные результаты показали, что инициируют диалог чаще всего репрезентативы, т.е. речевые акты, связанные с обменом информацией (38% случаев), «этикетное» начало (приветствия, вокативы) имеет место в 23% диалогов, а в 19% случаев разговор начинается с регулятивной формы. Речевые акты, завершающие диалог, показывают большее разнообразие: это репрезентативы (16% случаев), оценочные суждения (валюативы) (14%), регулятивные формы (14%), по 8% — директивы, комиссивы и этикетные формы и 7% — экспрессивы. Наиболее типичными бинарными последовательностями речевых актов оказались: два репрезентатива подряд (22,35%), регулятивная форма и следующий за ней репрезентатив (6,93%), репрезентатив и регулятивная форма (6,0%), валюатив и следующий за ним репрезентатив (5,21%), репрезентатив и оценочное суждение (4,77%), а также двусторонняя комбинация директива с репрезентативом (по 2,77%).

Ключевые слова: русская разговорная речь, повседневная речевая коммуникация, структура диалога, речевой акт, n-граммы, речевой корпус, устный дискурс

[1] Раздел 4 данной работы подготовлен при поддержке гранта РНФ «Система прагматических маркеров русской повседневной речи» (проект № 18-18-00242).

THE STRUCTURE OF EVERYDAY DIALOGUE AS THE SEQUENCE OF SPEECH ACTS

Sherstinova T. Yu. (t.sherstinova@spbu.ru)
Philological Faculty of Saint-Petersburg State University;
National Research University Higher School of Economics;
Saint-Petersburg, Russia

The structure of Russian everyday dialogue was studied on the basis of 73 microdialogues of everyday speech communication from the 'One Day of Speech' corpus (the ORD Corpus). The aim of the research was to find out what types of speech acts commonly initiate and complete everyday dialogues, as well as to reveal the most typical sequences of speech acts in these dialogues. Altogether, 2230 speech acts of 30 people referring to both professional, and household conversations have been analysed. N-gram analysis has been used to calculate the most frequent sequences of speech acts. The obtained results showed that dialogues are usually started by representatives, i. e. speech acts related to the exchange of information (38% of all cases), etiquette beginnings (greetings, vocatives) take place in 23% of the dialogues, and in 19% of cases the conversation begins with a regulative form. Speech acts ending dialogues show a greater variety: representatives contribute 2% of all dialogue ends, valuative judgments and regulatory forms cover 14% each, further go directives (8%), commissions (8%), etiquette forms (8%) and emotional and expressive form (7%). As for the most typical bigrams of speech acts, they are the following: two consecutive representatives (22.35%), a regulatory form followed by a representative (6.93%), a representative and a regulatory form (6%), a valuative with a following representative (5.21%), a representative and a valuative judgment (4.77%), as well as two combinations of a directive with a representative (2.77% each). Besides, the article presents data on the occurrence of the most frequent pairs of speech acts at the subtype level. Here, the most frequent one is the sequence 'question'+'answer', which covers 2.45%.

Key words: spoken Russian, everyday speech communication, dialogue structure, speech act, n-gram, speech corpus, oral discourse

1. Введение

Задача данного исследования состоит в выявлении наиболее типичных последовательностей речевых актов (РА) в структуре повседневного диалога. Материалом исследования являются 73 микродиалога повседневного общения из речевого корпуса «Один речевой день» (ОРД). Особенностью звукозаписей корпуса ОРД является их «аутентичность» — они выполняются в естественных условиях коммуникации с применением методики долговременной записи речи информанта в течение всего дня, «от восхода до заката» с помощью диктофона [Asinovsky et al. 2009]; [Bogdanova-Beglarian et al. 2016]. Такая методика

сбора речевого материала традиционно используется в японских лингвистических исследованиях (см., например, [Сибата 1983]; [Campbell 2004]), она также применялась при подготовке материалов для демографического подкорпуса Британского национального корпуса [Burnard 2007].

В настоящее время корпус ОРД содержит более 1250 часов звукозаписей речи, полученной от 130 информантов, мужчин и женщин, представителей разных профессий, разного социального статуса, в возрасте от 17 до 83 лет, и более тысячи их коммуникантов. Богатая коллекция звукозаписей корпуса дает возможность проводить на его материале лингвистические исследования повседневной речи в разных условиях коммуникации для разных языковых уровней (см. например, [Богданова-Бегларян и др. 2017]; [Bogdanova-Beglarian et al. 2016]; [Шерстинова 2016 и др.]).

Кроме того, коллекция ОРД дает уникальную возможность исследовать прагматику и структуру повседневных диалогов. Для прагматического анализа речевой коммуникации в корпусе ОРД была разработана многоуровневая система аннотирования коммуникативных эпизодов на макро- и микроуровнях[2]. Так, вся речевая продукция информанта в течение периода записи членится на *макроэпизоды*, объединенные локусом коммуникации, ее основной задачей и участниками [Sherstinova 2015]. Единицей хранения данных в корпусе являются макроэпизоды речевой коммуникации, записанные в виде отдельных звуковых файлов. Макроэпизоды далее сегментируются на единицы меньшего уровня, именуемые *микроэпизодами*, объединенные по теме разговора и частной коммуникативной задаче [Шерстинова 2015]. Исследование структуры диалога в настоящей работе проведено на материале микроэпизодов (микродиалогов), которые можно считать относительно однородными (по теме и задачам) единицами коммуникации.

2. Система аннотирования речевых актов в корпусе ОРД

Исходное понятие «речевого акта», изначально предложенное Дж. Остином [Austin 1962], затем существенно переработанное Дж. Сёрлем [Searle 1976], а потом и многими другими их последователями, к настоящему времени претерпело значимую транформацию. Сам термин «речевой акт» является достаточно популярным в современной лингвистике, однако определения РА и конкретные категории РА в трудах разных ученых и в разных научных школах сильно разнятся.

В прагматических исследованиях, проводимых на материале корпуса ОРД, под речевым актом понимается целенаправленное речевое действие, рассматриваемое в контексте прагматической ситуации и обладающее определенной иллокутивной силой. Предполагается, что любая реплика говорящего состоит из одного или нескольких РА. Более того, мы полагаем, что любое высказывание может быть интерпретировано как речевой акт определенного типа. В этом смысле наш подход отличается от традиционного понимания РА, предложенного

[2] Разработка системы прагматического аннотирования была поддержана грантом РГНФ № 12-04-12017 «Информационная система коммуникативных сценариев спонтанной русской речи».

Дж. Сёрлем. Более близким для нас является позиция М. М. Бахтина, который считал, что любое высказывание, каким бы кратким (и даже незавершенным) оно ни было, имеет некоторое значение или несет определенную информацию о говорящем [Bakhtin 1986]. Поэтому мы исходим из положения, что любая устная реплика может быть интерпретирована (проаннотирована) с позиции своего прагматического значения (роли) в процессе коммуникации. Формальное аннотирование речевых актов подразумевает их предварительное вычленение (сегментацию) и выполняется вручную в среде ELAN [Sloetjes & Wittenburg 2008].

При разработке классификационной схемы РА для аннотирования корпуса ОРД были проанализированы наиболее известные системы формального представления речевых действий, используемые в разных лингвистических корпусах с прагматической разметкой: например, корпус SPAAC (Speech Act Annotated Corpus, Великобритания) [Weisser 2003]; [Leech & Weisser 2003], система Dialogue Act Markup in Several Layers (DAMSL) [Allen & Core 1997], международный проект Cross-Cultural Study of Speech Act Realization Patterns [Blum-Kulka & Olshtain 1984], VRM-система дискурсивной таксономии (Verbal Response Modes), предложенная В. Стайлесом [Stiles 1992], и некоторые другие [Weisser 2014].

Однако большинство предлагаемых классификаций было разработано для ограниченного перечня коммуникативных сценариев (например, обращение в колл-центр по телефону или покупка железнодорожных билетов) и поэтому не подходит для аннотирования такого сложного жанра, как повседневное речевое общение. Для решения нашей задачи целесообразным показалось использовать классификации речевых поступков, разработанные отечественными лингвистами специально для русской разговорной речи [Борисова 2009].

Основные типы речевых актов, аннотируемых в корпусе ОРД, определяются следующим образом [Шерстинова 2015].

- *Репрезентативы* (ИНФ) — РА, главной целью которых является обмен информацией между участниками диалога.
- *Директивы* (ДИР) — побуждают адресата к действию (или бездействию) или выражают попытку повлиять на его мировоззрение, эмоции и установки.
- *Комиссивы* (КОМ) — связаны с принятием говорящим на себя определенных обязательств.
- *Экспрессивы-эмотивы* (ЭМО) — используются для выражения и передачи чувств и эмоций.
- *Этикетные экспрессивы* (ЭТИ) — стандартизованные формы, регулирующие коммуникацию в этикетных и ритуализированных ситуациях.
- *Валюативы* (ВАЛ) — используются для выражения оценочного мнения или мнения-суждения.
- *Суппозитивы* (СУП) — выражают мнение или предположение говорящего.
- *Коммуникативные регулятивы* (РЕГ) — фатические речевые поступки, связанные с «организационными» аспектами взаимодействия, используемые для структурирования и ведения диалога.

Особенностью реальной речевой коммуникации является тот факт, что часть высказываний остаются незавершенными, оборванными, «брошенными

на полуслове». При этом прозвучавшие фрагменты не всегда дают возможность реконструировать первоначальную интенцию говорящего, т. е. какой тип речевого акта планировался к воспроизведению. Такие сегменты речевого потока в корпусе помечаются кодом (ФРА), что означает *неопределяемый фрагмент* — т. е. незавершенный РА, по которому невозможно определить его иллокутивную силу.

Далее, поскольку записи ОРД выполняются в «полевых» условиях, где естественным является наличие фоновых шумов (шум улицы, посторонние разговоры, звук телевизора и др.), то некоторые речевые фрагменты оказываются *нерасшифруемыми* (НЕР). Так помечаются РА, для которых невозможно получить текстовую расшифровку вследствие особенностей коммуникации или качества звукозаписи.

Наконец, в устной речевой коммуникации не последнюю роль играют *паралингвистические события* (ПАР), многие из которых могут иметь иллокутивную силу (например, смех, вздох, стон и пр.).

В мультивекторных разговорах (когда имеют место несколько параллельных разговоров) аннотируется только основная линия коммуникации, посторонние и параллельные реплики и диалоги не сегментируются на речевые акты и соответствующим образом не аннотируются.

В каждом основном типе речевых актов выделяются подтипы, необходимые для того, чтобы, например, отделить *вопросы* от *ответов* в общей категории *репрезентативов* или *просьбу* от *приказа* среди директивов. Некоторые из наиболее частотных подтипов РА для каждого основного типа приведены в табл. 1.

Таблица 1. Некоторые подтипы РА, выделяемые при аннотировании

№	Основной тип РА	Подтипы РА
1	Репрезентатив	Дескриптив, констатив, сообщение, экспликатив, комментарий, вопрос (рогатив), напоминание, уточнение, репродуктив (передача чужой речи) и др.
2	Регулятив	Тематическая инициатива, речевая поддержка, речевая придержка, запрос на наличие контакта, отказ от темы, смена темы, переспрос, показатель готовности к коммуникации и др.
3	Валюатив	Возражение, отрицание, опровержение, согласие и т. п.
4	Директив	Просьба, требование, приказ, совет, убеждение, разрешение, утешение и др.
5	Этикетная форма	Вокатив, приветствие, прощание, извинение, благодарность, пожелание, поздравление и др.
6	Экспрессив-эмотив	Выражение положительных эмоций, выражение отрицательных эмоций, удивление и т. п.
7	Комиссив	Обещание, согласие выполнить просьбу, заявление о намерениях, отклонение предложения и др.
8	Суппозитив	Предположение, выражение личного мнения и под.
9	Паралингвистическое событие	Смех, вскрик, вздох, стон, свист, цоканье языком и другие невербальные события, значимые для процесса коммуникации.

При аннотировании РА в корпусе ОРД разметка выполняется на следующих четырех уровнях:

1. RAct (речевой акт) — «форма выражения» речевого акта, т. е. орфографическая запись соответствующего фрагмента речи с использованием синтагматической и фразовой разметки.
2. RActSp (говорящий) — код информанта или коммуниканта.
3. RAGenT — общий тип речевого акта.
4. RADetT — детализация речевого акта (подтип).

В табл. 2 приведен пример аннотирования фрагмента диалога с использованием разработанной схемы.

Таблица 2. Пример аннотирования фрагмента диалога

Реплика	Код говорящего	Речевой акт	Тип РА	Подтип РА
Р1	Ж1	*так //*	РЕГ (регулятив)	ИНИ (начало темы)
	Ж1	*всё-таки не туда ? <пауза> или туда?*	ИНФ (репрезентатив)	ВОПР (вопрос)
Р2	М1	*мне кажется / что вон туда // до конца //*	СУП (суппозитив)	ПРЕДП (предположение)
Р3	Ж1	*это не она //*	ВАЛ (валюатив)	ВОЗР (возражение)
	Ж1	*вот туда //*	ИНФ (репрезентатив)	КОРР (коррекция)
Р4	М1	*но там зато фонарь виден хорошо //*	ИНФ (репрезентатив)	ЭКСП (экспликатив)
Р5	Ж1	*угу //*	ВАЛ (валюатив)	СОГЛ1 (согласие)

3. Материал и методика исследования

Как уже было отмечено во введении, материалом настоящего исследования стали 73 микроэпизода повседневного общения, включающие как рабочие, так и домашние разговоры и общение друзей. Анализируемые диалоги относятся к 6 макроэпизодам и включают в общей сложности речь 30 человек — 6 информантов и 24 их коммуникантов, мужчин и женщин, представителей разных возрастных и профессиональных групп.

Весь речевой материал был поделен на 2230 речевых актов, которые были вручную проаннотированы. Полученная статистика о распределении основных типов РА в исследуемой выборке приведена в табл. 3.

Таблица 3. Распределение основных типов речевых актов

№	Тип РА	Аббревиатура	Абс. кол-во	%
1	Репрезантатив	ИНФ	884	39,62
2	Регулятив	РЕГ	279	12,51
3	Валюатив	ВАЛ	254	11,39
4	Директив	ДИР	151	6,77
5	Этикетная форма	ЭТИ	93	4,17
6	Паралингвистическое событие	ПАР	83	3,72
7	Экспрессив-эмотив	ЭМО	79	3,54
8	Комиссив	КОМ	62	2,78
9	Суппозитив	СУП	58	2,60
10	Незавершенный фрагмент	ФРА	57	2,55
11	Нерасшифруемый фрагмент	НЕР	18	0,81
12	Смешанные типы	ИНФ/ЭМО, ИНФ/РЕГ, ИНФ/ВАЛ, ДИР/ЭТИ и др.	249	11,16

Таким образом, почти 40% всех высказываний относятся к категории репрезентативов, главная задача которых состоит в обмене информацией, 12,5% речевых актов являются регулятивными формами, задающими и поддерживающими ход диалога, 11,4% высказываний являются валюативами, выражающими оценочные мнения или суждения. Значительно реже встречаются директивы (6,7%), этикетные формы (4,2%) и др. категории РА, а 9,38% всех высказываний выборки представляют собой смешанные типы.

Следует обратить внимание на то, что паралингвистические явления составляют довольно значимый компонент устной коммуникации (3,7%). Подробнее о частоте паралингвистических событий в повседневной речи см. [Sherstinova 2018].

Рассмотрим, какие подкатегории речевых актов оказались самыми частотными для основных типов (см. табл. 4). Процент посчитан по отношению ко всем речевым актам данного типа.

Таким образом, среди репрезентативов самые частые категории — это *вопросы* (28,65%) и *экспликативы* (*разъяснения*, 17,35%), наиболее употребительные регулятивные формы — это разного рода *маркеры границ* (начала, сегмента, конца — *так, вот, ну* и т.п.), *речевая поддержка* (*ага*, *угу*, *да* и т.п.), *переспрос*. Среди валюативов на нашей выборке больше всего встретилось выражения *согласия* или *одобрения* (37,84%), в то время как *возражение* имело место почти в 4 раза реже (9,96%). Наиболее частотные директивы — *предложение* (25,83%) и *просьба* (19,87%). Среди этикетных форм выделяются *вокативы* (треть всех реализаций) и *приветствия* (28%), а среди эмотивов — *положительные эмоции* (21%) и *удивление* (14%). Коммисивы и суппозитивы в речевой коммуникации используются нечасто и их подтипы немногочисленны: для коммисивов это *согласие выполнить просьбу* (30%), *заявление о намерениях* (30%) и *обещание* (12%), а для суппозитивов — *предположение* (65%) и *выражение личного мнения* (23%).

Таблица 4. Наиболее частотные подтипы речевых актов в исследуемой выборке

№	Тип РА	Подтип	Абс. кол-во	%
1	Репрезантатив	Вопрос (ВОПР)	251	28,65
		Экспликатив (ЭКСП)	152	17,35
		Сообщение (СООБ)	119	13,58
		Ответ (ОТВ)	105	11,98
2	Регулятив	Маркер начала (ИНИ)	55	19,78
		Речевая поддержка (РП)	41	14,74
		Переспрос (ПЕРЕ)	34	12,23
		Маркер границы (ВОТ)	25	8,99
		Подтверждение понимания	24	8,63
3	Валюатив	Согласие, одобрение (СОГЛ)	95	37,84
		Суждение (СУЖ)	31	12,35
		Возражение (ВОЗ)	25	9,96
4	Директив	Предложение (ПРЕД)	39	25,83
		Просьба (ПРОС)	30	19,87
		Привлечение внимания (ВНИМ)	22	14,57
		Инструкция (ИНС)	9	5,96
5	Этикетная форма	Вокатив (ВОК)	33	35,48
		Приветствие (ПРИВ)	26	27,95
		Прощание (ПРОЩ)	5	5,37
6	Паралингвистическое событие	Смех	57	71,25
		Вздох	10	12,50
7	Экспрессив-эмотив	Положительные эмоции (ЭПОЛ)	16	20,78
		Удивление (УДИВ)	11	14,28
		Реакция на внешний источник (РЕАК)	10	12,99
		Недовольство (НЕД)	9	11,68
		Отрицательные эмоции (ЭОТР)	7	9,09
8	Комиссив	Согласие выполнить просьбу и под. (КСОГЛ)	18	30,51
		Заявление о намерениях (ЗАЯН)	18	30,51
		Обещание (ОБЕ)	7	11,86
9	Суппозитив	Предположение (ПРЕДП)	37	64,91
		Выражение личного мнения (МНЕ)	13	22,81
10	Смешанные типы	Вопрос+выражение эмоций Вопрос+начало темы Ответ+выражение эмоций Просьба+вопрос Просьба+выражение эмоций и др.	249	11,16*

* в данном случае — от общего кол-ва РА

Для подсчета наиболее частотных последовательностей РА использовалась методика n-граммного анализа — довольно популярный статистический метод, позволяющий строить вероятностные лингвистические модели. Следует отметить, что в данном исследовании диалог рассматривается как единая структура, состоящая из последовательности речевых актов; длительность пауз между репликами и смена коммуникативных ходов говорящими при этом не учитываются[3].

4. Наиболее типичные бинарные последовательности речевых актов и частотность отдельных типов речевых актов в начале и конце диалога

В табл. 5 приведены данные о частоте встречаемости пар РА по основным типам. В первом столбце приведены начальные элементы биграмм, в первой строке — соответствующие им вторые элементы. Доля встречаемости представлена в процентах. Начало и конец микродиалогов обозначены в таблице как (НД) и (КД), что позволяет определить речевые акты, преобладающие в начале и конце диалога. Наиболее частотные биграммы РА в таблице маркированы.

Таблица 5. Доля встречаемости пар речевых актов в исследуемой выборке (%)

	ЭТИ	ЭМО	СУП	РЕГ	ПАР	КОМ	ИНФ	ДИР	ВАЛ	(КД)	Итого
(НД)	0,33	0,28	0,17	0,55	0,22	0,33	0,67	0,33	0,55	—	3,44
ВАЛ	0,17	0,28	0,61	1,33	1,11	0,39	**5,21**	1,05	2,55	0,06	**12,76**
ДИР	0,55	0,33	0,33	1,22	0,11	0,28	2,77	0,78	0,94	0,28	**7,60**
ИНФ	1,44	1,16	1,44	**6,05**	1,33	0,89	**22,35**	2,77	**4,77**	1,11	**43,32**
КОМ	0,11	0,11	0,00	1,11	0,00	0,39	0,78	0,55	0,33	0,00	3,38
ПАР	0,11	0,33	0,06	0,06	0,17	0,00	1,66	0,39	0,89	0,06	3,72
РЕГ	0,44	0,44	0,22	2,38	0,33	0,67	**6,93**	0,94	1,28	0,78	**14,42**
СУП	0,06	0,00	0,17	0,55	0,11	0,00	1,16	0,22	0,50	0,00	2,77
ЭМО	0,17	0,72	0,06	0,33	0,22	0,00	1,50	0,17	0,44	0,11	3,72
ЭТИ	1,22	0,17	0,00	0,50	0,11	0,22	0,83	0,28	0,61	0,94	**4,88**

Из таблицы 5 видно, что наиболее типичными бинарными последовательностями РА оказались: два репрезентатива подряд (ИНФ+ИНФ, 22,35%), регулятивная форма и следующий за ней репрезентатив (РЕГ+ИНФ, 6,93%), репрезентатив и регулятивная форма (ИНФ+РЕГ, 6,05%), валюатив и следующий за ним репрезентатив (ВАЛ+ИНФ, 5,21%), репрезентатив и оценочное суждение (ИНФ+ВАЛ, 4,77%), а также двусторонняя комбинация директива с репрезентативом (ДИР+ИНФ и ИНФ+ДИР — по 2,77%).

[3] Тем не менее, представляется целесообразным рассмотреть эти факторы при продолжении прагматических исследований корпуса ОРД в будущем.

Из таблицы можно получить и информацию об относительной частотности отдельных типов речевых актов в начале и конце диалога. Однако для удобства интерпретации данных их процент лучше пересчитать по отношению ко всем позициям начала-конца микроэпизода.

В итоге оказалось, что *инициируют диалог* чаще всего репрезентативы, т. е. РА, связанные с обменом информацией (38 % случаев от всех начальных позиций микродиалогов), «этикетное» начало (приветствия, вокативы) имеет место в 23 % диалогов, а в 19 % случаев разговор начинается с регулятивной формы.

Речевые акты, завершающие диалог, показывают большее разнообразие: это репрезентативы (16 %), оценочные суждения (валюативы) (14 %), регулятивные формы (14 %), по 8 % — директивы, комиссивы и этикетные формы и 7 % — экспрессивы.

Однако данные, приведенные в табл. 5, могут показаться недостаточно информативными, если мы хотим проанализировать структуру диалога на более детальном уровне. Воспользуемся для этой цели частотным списком биграмм, полученных при анализе речевых актов на уровне подтипов (см. табл. 6).

Таблица 6. Наиболее частотные пары речевых актов на уровне подтипов

Ранг	Последовательность	%
1	*вопрос — ответ*	2,45
2	*вопрос — вопрос*	1,01
3	*экспликатив — экспликатив*	0,97
4	*ответ — вопрос*	0,93
5	*экспликатив — вопрос*	0,93
6	*сообщение — сообщение*	0,89
7	*вопрос — экспликатив*	0,76
8	*экспликатив — согласие*	0,55
9	*согласие — вопрос*	0,51
10	*сообщение — вопрос*	0,46

Видно, что все наиболее частотные пары РА относятся к категории репрезентативов. И, вполне предсказуемо, на первом месте здесь находится пара *вопрос — ответ*, покрывающая 2,45 % всех случаев.

Вопрос оказался также весьма типичным началом разговора: с вопроса начинаются 16 % всех микродиалогов исследованной выборки. В 17 % случаев микродиалог инициируется маркером начала новой темы (например, репликой *так*). Также диалоги регулярно начинаются с приветствия (13 %), с сообщения (11 %) или с вокатива (10 %).

Что касается реплик, завершающих диалог, на уровне подтипов РА почти не оказалось регулярно повторяющихся. Самым частотным финалом диалога оказалось *согласие*, который завершает диалог в 10 % случаев, относительно частотны *сообщения* и *регулятивные формы* (напр., *вот*) (по 4 % случаев). Таким образом, начало диалога кажется намного проще для формального моделирования структуры диалога, чем его конец.

5. Заключение

Проведенное исследование показало, что инициируют диалог чаще всего репрезентативы, т. е. речевые акты, связанные с обменом информацией (38% случаев), «этикетное» начало (приветствия, вокативы) имеет место в 23% диалогов, а в 19% случаев разговор начинается с регулятивной формы. РА, завершающие диалог, показывают большее разнообразие: это репрезентативы (16%), оценочные суждения (валюативы) (14%), регулятивные формы (14%), по 8% — директивы, комиссивы и этикетные формы и 7% — экспрессивы. Наиболее типичными бинарными последовательностями РА оказались: два репрезентатива (17,5%), репрезентатив и регулятивная форма (5,4%), регулятивная форма и следующий за ней репрезентатив (4,7%), репрезентатив и оценочное суждение (4,1%), валюатив и следующий за ним репрезентатив (3,7%), а также двусторонняя комбинация директива с репрезентативом (по 2,2%).

Разумеется, при интерпретации любых корпусных результатов большое значение имеет используемая в корпусе система аннотирования данных. Схема аннотирования речевых актов, разработанная для корпуса ОРД, как и многие другие формальные классификации, предназначенные для анализа такой сложной и многоаспектной материи, как устная речь, конечно, не является идеальной и лишенной недостатков. Так, не всегда атрибуцию РА по типу и подтипу можно произвести однозначно и непротиворечиво. Однако на сегодняшний день это оптимальная структура, к которой мы пришли в результате анализа реального речевого материала, и которая, на наш взгляд, в большинстве случаев дает вполне корректный и однозначно интерпретируемый результат.

Мне бы хотелось поблагодарить анононимных рецензентов данной работы за их конструктивные замечания и пожелания, большинство из которых обязательно будут приняты во внимание при продолжении исследования. Наиболее интересным мне кажется предложение размечать речевые акты по нескольким независимым параметрам, выделив те или иные аспекты субъективной модальности в независимое поле описания. Развивая эту идею, можно было бы также маркировать и эмоциональную окраску РА, и наличие в нем регулирующих разговор прагматических маркеров — то есть те факторы, которые, как видно из табл. 3, чаще других приводят к появлению «смешанных типов» в используемой эмпирической классификации. При разработке такой системы аннотирования, по-видимому, следует еще раз обратить внимание на VRM-систему дискурсивной таксономии В. Стайлеса [Stiles 1992], в которой подобный подход уже был реализован. Многоаспектное аннотирование речевых актов представляется перспективным, хотя достаточно трудоемким, и приведет к необходимости полного пересмотра используемой классификации (и, следовательно, повторного аннотирования речевого материала).

Надо сказать, что в прагматических исследованиях существует представление о том, что сколько в языке имеется глаголов, связанных с описанием коммуникации, столько разнообразных подкатегорий речевых актов говорящие способны выделять. Действительно, предлагаемые исследователями списки категорий довольно представительны и насчитывают порядка 200 единиц. Но так как многие из них, с одной стороны, близки по значению

или синонимичны, а с другой — не учитывают «реального наполнения» прагматического компонента высказывания (который, как выясняется, до сих пор не имеет адекватного «словарного» описания), построение единой и непротиворечивой классификации РА, которая служила бы основой для их корректной лингвистической и прагматической интерпретации, можно считать одной из «гильбертовых проблем» современной лингвистики.

Литература

1. *Богданова-Бегларян Н. В., Блинова О. В., Мартыненко Г. Я., Шерстинова Т. Ю.* (2017), Некоторые инвариантные характеристики русской разговорной речи: фонетика, морфология, синтаксис // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 31 мая — 3 июня 2017 г.). Вып. 16 (23): В 2 т. Т. 2 — М.: Изд-во РГГУ, 2017, с. 75–86.
2. *Борисова И. Н.* (2009), Русский разговорный диалог. Структура и динамика. М.: Книжный дом «ЛИБРОКОМ», 320 с.
3. *Сибата Т.* (1983), Исследования языкового существования в течение 24 часов // Алпатов В. М., Вардуль И. Ф. (ред.) Языкознание в Японии, М.: Радуга, с. 134–141.
4. *Шерстинова Т. Ю.* (2015), Прагматическое аннотирование коммуникативных единиц в корпусе ОРД: микроэпизоды и речевые акты / Захаров В. П., Хохлова М. В. (ред.) Труды международной конференции «КОРПУСНАЯ ЛИНГВИСТИКА — 2015». СПб.: СПбГУ, с. 436–445.
5. *Шерстинова Т. Ю.* (2016), Наиболее употребительные слова повседневной русской речи (в гендерном аспекте и в зависимости от условий коммуникации) // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Москва, 1–4 июня 2016 г.). Вып. 15 (22). М.: Изд-во РГГУ, 2016, с. 616–631.
6. *Allen J. & Core M.* (1997), Draft of DAMSL: Dialog act markup in several layers. Online at: https://www.cs.rochester.edu/research/speech/damsl/RevisedManual/ (last accessed on Feb. 17, 2018).
7. *Asinovsky A., Bogdanova N., Rusakova M., Ryko A., Stepanova S., Sherstinova T.* (2009), The ORD Speech Corpus of Russian Everyday Communication «One Speaker's Day»: Creation Principles and Annotation. In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNAI, vol. 5729. Springer, Berlin-Heidelberg, pp. 250–257.
8. *Austin J. L.* (1962), How To Do Things With Words, Oxford University Press, Oxford.
9. *Bakhtin M. M.* (1986), Speech Genres and Other Late Essays, Univ. of Texas Press, Austin, TX.
10. *Bogdanova-Beglarian N., Sherstinova T., Blinova O., Ermolova O., Baeva E., Martynenko G., Ryko A.* (2016), Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech / Ronzhin, A. et al. (eds.) SPECOM 2016, Lecture Notes in Artificial Intelligence, LNAI, vol. 9811. Springer, Switzerland, pp. 659–666.

11. *Bogdanova-Beglarian N., Sherstinova T., Blinova O., Martynenko G.* (2017), Linguistic Features and Sociolinguistic Variability in Everyday Spoken Russian. In: Karpov A., Potapova R., Mporas I. (eds) Speech and Computer. SPECOM 2017. Lecture Notes in Computer Science, vol 10458. Springer, Cham, pp. 503–511.
12. *Blum-Kulka S. & Olshtain E.* (1984), Requests and Apologies: A Cross-Cultural Study of Speech Act Realization Patterns (CCSARP) Applied Linguistics, Vol. 5, No. 3, pp. 196–213.
13. *Burnard L.* (Ed.) (2016), Reference guide for the British National Corpus (XML edition). Published for the British National Corpus Consortium by Oxford University Computing Services, 2007. Available at: <http://www.natcorp.ox.ac.uk/docs/URG/>. Retrieved: February 20, 2018.
14. *Leech G. & Weisser M.* (2003), Generic speech act annotation for task-oriented dialogues, Proc. of the Corpus Linguistics 2003 Conference. Lancaster University: UCREL Technical Papers, V. 16.
15. *Searle J. R.* (1976), A classification of illocutionary acts, Language in Society, 5(1), pp. 1–23.
16. *Sherstinova T.* (2018), Audible Paralinguistic Phenomena in Everyday Spoken Conversations: Evidence from the ORD Corpus Data, Language, Music, and Computing (LMAC-2017), Communications in Computer and Information Science, CCIS, Springer International Publishing Switzerland (in print)
17. *Sloetjes H. & Wittenburg P.* (2008), Annotation by category — ELAN and ISO DCR, Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008.
18. *Stiles W.* (1992), Describing Talk: A Taxonomy of Verbal Response Modes, Sage, Newbury Park, CA.
19. *Weisser M.* (2003), SPAACy: A semi-automated tool for annotating dialogue acts. International Journal of Corpus Linguistics 8 (1), pp. 63–74
20. *Weisser M.* (2014), Speech act annotation, Corpus Pragmatics: a Handbook, CUP, Cambridge, pp. 84–111.

References

1. *Allen, J. and Core, M.* (1997), Draft of DAMSL: Dialog act markup in several layers. Online at: https://www.cs.rochester.edu/research/speech/damsl/RevisedManual/ (last accessed on Feb. 17, 2018).
2. *Asinovsky A., Bogdanova N., Rusakova M., Ryko A., Stepanova S., Sherstinova T.* (2009), The ORD Speech Corpus of Russian Everyday Communication «One Speaker's Day»: Creation Principles and Annotation, TSD 2009, LNAI, vol. 5729, pp. 250–257.
3. *Austin J. L.* (1962), How To Do Things With Words, Oxford University Press, Oxford.
4. *Bakhtin M. M.* (1986), Speech Genres and Other Late Essays, Univ. of Texas Press, Austin, TX.

5. *Blum-Kulka S. and Olshtain E.* (1984), Requests and Apologies: A Cross-Cultural Study of Speech Act Realization Patterns (CCSARP) Applied Linguistics, Vol. 5, No. 3, pp. 196–213.
6. *Bogdanova-Beglarian N., Sherstinova T., Blinova O., Ermolova O., Baeva E., Martynenko G., Ryko A.* (2016) Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech / Ronzhin, A. et al. (eds.) SPECOM 2016, Lecture Notes in Artificial Intelligence, LNAI, vol. 9811. Springer, Switzerland, 2016, pp. 659–666.
7. *Bogdanova-Beglarian N., Sherstinova T., Blinova O., Martynenko G.* (2017), Linguistic Features and Sociolinguistic Variability in Everyday Spoken Russian, SPECOM 2017, LNCS, vol. 10458, pp. 503–511.
8. *Bogdanova-Beglarian N., Sherstinova T., Blinova O., Martynenko G.* (2017), Some invariant features of Russian everyday speech: Phonology, morphology, syntax, Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, 2(16), pp. 82–95.
9. *Borisova I. N.* (2009), Russian spoken dialogue. Structure and Dynamics [Russkiy razgovornyy dialog. Structura i dinamika], LIBROKOM [Knizhnyj dom LIBROKOM], Moscow.
10. *Burnard, L.* (Ed.) (2016), Reference guide for the British National Corpus (XML edition). Published for the British National Corpus Consortium by Oxford University Computing Services, 2007. Available at: <http://www.natcorp.ox.ac.uk/docs/URG/>. Retrieved: February 20, 2018.
11. *Leech G. & Weisser M.* (2003), Generic speech act annotation for task-oriented dialogues, Proc. of the Corpus Linguistics 2003 Conference. Lancaster University: UCREL Technical Papers, V. 16.
12. *Searle J. R.* (1976), A classification of illocutionary acts, Language in Society, 5(1), pp. 1–23.
13. *Sherstinova T.* (2015), Approaches to Pragmatic Annotation in the ORD Corpus: Microepisodes and Speech Acts [Pragmaticheskoe annotirovanie konnunicativnykh jedinic v korpuse ORD: mikroepisody i rechevye akty], Proc. of the Int. Conf. "Corpus linguistics-2015" [Trudy mezhdunarodnoy nauchnoy konferentsii "Korpusnaya linguistica-2013"], pp. 436–446.
14. *Sherstinova T.* (2015), Macro Episodes of Russian Everyday Oral Communication: towards Pragmatic Annotation of the ORD Speech Corpus, SPECOM 2015, LNAI, vol. 9319, pp. 268–276
15. *Sherstinova T.* (2016), The most frequent words in everyday spoken Russian (in the gender dimention and depending on communication settings), Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, pp. 616–631.
16. *Sherstinova T.* (2018), Audible Paralinguistic Phenomena in Everyday Spoken Conversations: Evidence from the ORD Corpus Data, Language, Music, and Computing (LMAC-2017), Communications in Computer and Information Science, CCIS, Springer International Publishing Switzerland (in print).
17. *Sibata, T.* (1983), Issledovanija jazykovogo sushhestvovanija v techenie 24 chasov [Studying the language life with the method of the 24 hour survey]. In: Alpatov, V. M., Vardul, I. F. (eds.) Jazykoznanie v Japonii [Linguistics in Japan], Raduga, Moscow, pp. 134–141.

18. *Sloetjes H. & Wittenburg P.* (2008), Annotation by category — ELAN and ISO DCR, Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008.
19. *Stiles W.* (1992), Describing Talk: A Taxonomy of Verbal Response Modes, Sage, Newbury Park, CA.
20. *Weisser M.* (2003), SPAACy: A semi-automated tool for annotating dialogue acts. International Journal of Corpus Linguistics 8 (1), pp. 63–74.
21. *Weisser M.* (2014), Speech act annotation, Corpus Pragmatics: a Handbook, CUP, Cambridge, pp. 84–111.
22. *Wittenburg P., Brugman H., Russel A., Klassmann A., Sloetjes H.* (2006) ELAN: a Professional Framework for Multimodality Research, Proc. of LREC 2006, Fifth International Conference on Language Resources and Evaluation.

IMPROVING TOPIC MODELS WITH SEGMENTAL STRUCTURE OF TEXTS

Skachkov N. A. (nikolaj-skachkov@yandex.ru)
Lomonosov Moscow State University

Vorontsov K. V. (voron@forecsys.ru)
Dorodnicyn Computing Centre of RAS,
Moscow Institute of Physics and Technology, Russia

Probabilistic topic modeling is a powerful tool of text analysis, that reveals topics as distributions over words and then softly assigns documents to the topics. Even though the aggregated distributions can be good with basic models, a sequential topic representation of each document is often unsatisfactory. This work introduces a method that allows to increase the quality of topical representation of each single text using its segmental structure. Our approach is based on Additive Regularization of Topic Models (ARTM), which is a technique for imposing additional criteria into the model. The proposed method efficiently avoids a bag-of-words assumption by considering the topical connections of words that co-occur in a local segment. We assume, that sequential sentences are topically and semantically coherent, while the number of topics in each particular text fragment is low. We apply our model to topic segmentation task and achieve a better quality than the current state-of-the-art TopicTiling algorithm. In further experiments we demonstrate that the proposed technique reveals an interpretable sequential structure of documents, while keeping a number of topics low, i.e. the sparsity of the model increases. Apart from topic segmentation, the constructed topical text embeddings can be used in any other applications, where the analysis of the document structure is desirable.

Keywords: Topic modeling, text segmentation, topic segmentation, topical embeddings, sparse embeddings, EM-algorithm

ИСПОЛЬЗОВАНИЕ СЕГМЕНТНОЙ СТРУКТУРЫ ДОКУМЕНТОВ ДЛЯ ПОСТРОЕНИЯ ТЕМАТИЧЕСКОЙ МОДЕЛИ

Скачков Н. А. (nikolaj-skachkov@yandex.ru)
Московский Государственный Университет
им. М. В. Ломоносова

Воронцов К. В. (voron@forecsys.ru)
Вычислительный центр им. А. А. Дородницына РАН,
Московский Физико-Технический Институт, Россия

1. Introduction

Topic modeling is a rapidly developing branch of statistical text analysis. Topic model uncovers a hidden semantic structure of the text collection and finds a highly compressed representation of each document by a set of its topics. From the statistical point of view, each topic is a set of words or phrases that frequently co-occur in many documents. The topical representation of a document captures the most important information about its semantics and therefore it is useful for many applications including information retrieval, classification, categorization, summarization and segmentation of texts [Vorontsov, 2014].

Despite many advantages, topic models are known to fail modeling the structure of the text inside the documents. Usually all the topics that are presented in the document are evenly distributed along the text. That is strongly connected with the bag of words assumption in the modeling of the texts. This assumption significantly simplifies the theoretical inference that allows to receive an iterative solution known as EM algorithm. But in many tasks, such as analysis of large documents, intra-document search, or dialog systems, it is important to model intra-document topic behavior with a good granularity.

One of the most popular topic models is Latent Dirichlet Allocation (LDA) proposed in [Blei, 2003]. LDA is a two-level Bayesian generative model, which assumes that topic distributions over words and document distributions over topics are generated from prior Dirichlet distributions.

Many authors successfully tried changing LDA generative model in such a way that some assumptions about text structure are incorporated.

For example, in [Balikas, 2016] a model called senLDA was built. In this model, all words in a sentence could have only one and the same topic label. In the experiments, this model converged faster than LDA, and the representation of documents provided by this model successfully complemented LDA representation for a task of document classification.

In [Du, 2013] a more complex LDA-based Topic Segmentation Model (TSM) was proposed. It assumes that documents consist of segments, whose topical subjects are also present in the document subjects. To model the segments subjects, a Pitman-Yor process is used. It represents each segment as a Chinese restaurant, where customers represent words, dishes represent topics and tables represent monothematic subsets of words.

For all these models, any assumptions on the text structure change the generative model, thus making it hard to design and infer new modifications.

In this work, we offer a new method based on Additive Regularization of Topic Models (ARTM) [Vorontsov, 2014]. Our method allows to reconstruct the segmental structure of the text in the topic model. The estimated segment boundaries are being used to reduce the number of topics within a segment. This is made in an assumption that words within a text fragment share the same small set of topics. The result topic model granulates topics in the segments and increases the model sparsity. This topic structure may be used for automatic intra-document analysis.

Finally the proposed enhancements don't complicate the structure of model training and theoretical inference, but increase topic model quality, sparsity and interpretability. All of that reveals new opportunities for applications of topic models.

To evaluate the segmentation quality an artificially generated corpus is used. It is generated from PostScience collection. We are using artificial documents for evaluation as it was done in many works before [Galley, 2003; Du, 2013; Riedl, 2012]. We use them because the comparison on real texts is a complicated challenge as there are no golden standard segment boundaries provided. We create artificial documents by concatenating full-source documents from the PostScience collection. This method was shown to be more justified comparing to the other ways of artificial documents construction [Riedl, 2012].

The paper is organized as follows. The next section gives an overview of ARTM and section 3 introduces our approach to topic segmentation. Section 4 provides details about parameter evaluation. Finally, section 5 presents our final results on an artificial corpus.

2. Additive regularization of topic models

A topic model describes a collection D by a finite set of topics T. In ARTM [Vorontsov, 2014] and in more basic PLSA model [Hoffman, 1999], the distribution of words in documents is modeled as a mixture of topics:

$$p(w|d) = \sum_{t\in T} p\,(w|t)p(t|d), d \in D, w \in W \tag{1}$$

The model is parametrized stochastic matrices Φ and Θ with the elements:

$$\varphi_{wt} = p(w|t), \theta_{td} = p(t|d)$$

Topic modeling can be also interpreted as a task of approximate matrix factorization $F \approx \Phi\Theta$. The solution of matrix factorization task is non-unique, thus we follow ARTM [Vorontsov, 2012] approach and consider additional criteria to learn better Φ and Θ matrices. Particularly, we maximize the weighted sum of log-likelihood and some additive regularizers R_i:

$$\sum_{d\in D}\sum_{w\in W} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + \sum \tau_i R_i(\Phi, \Theta) \to \max_{\Phi,\Theta} \tag{2}$$

Regularizers R_i impose additional problem-specific criteria on the model parameters. Regularizer coefficients τ_i balance the importance of regularizers and log-likelihood. If no regularizers are added, the described model simplifies to PLSA.

The stationary point of the problem (2) satisfies the system of equations, that yields expectation-maximization algorithm as the fixed point iteration method. E-step of this algorithm calculates probabilities of word assignments to topics in the context of a document $(t \mid d, w) \equiv p_{tdw}$. M-step uses these probabilities to update matrices Φ and Θ.

3. Using the segmental structure of documents to improve EM-algorithm

According to (1), each document is represented as a bag-of-words. Additive regularizers are normally applied on M-step and they also cannot make any assumptions about the word order.

However, during the E-step, we compute p_{tdw} probabilities for each position in the document sequentially. It means, we can impose additional assumptions on topic distributions for the words that occur in the same part of the text. According to these assumptions, p_{tdw} values can be modified and then used at the M-step of the EM algorithm.

In real texts, authors usually convey their thoughts in a sequential way. That is why we can expect to see only a few topics in any small piece of text. This can be formulated as a sparsity assumption of subjects within a sentence ($t|s$), where we define the subject of the sentence as an average sum of its word distributions over topics:

$$p_{ts} \equiv p(t|s) = \frac{1}{n_s} \sum n_{sw}\, p_{tdw},$$

where n_s is the length of sentence, and n_{sw} is the number of occurrences of word in sentence.

One can show that the provided sparsity assumption influences p_{tdw} in the following way:

$$p_{\widetilde{tdw}} = p_{tdw}\left(1 - \frac{\tau}{n_{dw}} \sum_{s \in S_d} \frac{n_{sw}}{n_s}\left(\frac{1}{p_{ts}} - \sum_{z \in T} \frac{p_{zdw}}{p_{zs}}\right)\right) \tag{3}$$

where S_d is a set of all sentences in a document.

We omit derivation of this formula due to space limitations, but provide some intuition behind. If p_{ts} for some sentence is close to zero, then $1/p_{ts}$ is big and the resulting sign for this sentence term is negative. It means, that the probability of the corresponding topic in the word will be decreased. On the contrary, if p_{ts} is close to 1, this sentence term may be positive and the probability of the corresponding topic will increase. Shortly speaking, each term of the sum in this formula brings the distributions p_{tdw} closer to the main topics of the sentences where the word occurs.

It is worth being mentioned that the $p_{\widetilde{tdw}}$ may not determine a distribution over topics, because the values can be negative, but this doesn't break the EM algorithm.

The τ parameter in the formula determines the strength of the influence of sentence subjects on word distributions over topics.

Now let us elaborate the idea of sparsity of the subjects of text fragments and use it for topic segmentation task. We assume that any text consists of *segments* that can be represented by a small number of topics. Topics are supposed to stay constant along each segment. Two sequential segments are supposed to have low intersection of topics.

We estimate borders of the segments gradually while the model is being learnt. At the first iteration of EM-algorithm we use sentences as an initial approximation of the segments. Then at each E-step we find the subjects of the segments and merge the sequential segments if they share the same topics. Equation (3) in this method is applied to the segments that have been built so far by the current iteration.

4. Topic model quality evaluation

To evaluate topic model quality, we will take into account two factors: topic model segmentation quality and intra-document sparsity. Segmentation quality shows the ability of the model to restore topic borders and semantic changes in the text. To evaluate this, we will use an artificially generated text corpus. It will provide us with the gold standard for segment boundaries in the documents. To check how the golden standard boundaries overlap with the estimated ones, we will use P_k and WindowDiff measures as it was done in the prior work [Riedl, 2012].

The intra-document sparsity shows the ability of the model to describe semantic segments with the smallest possible number of topics. The sparsity of segment subjects implies the sparsity of the whole document, so we will use Θ-matrix average sparsity to evaluate this.

4.1. Building segment boundaries using topic models

For all topic models, we will use a special topic segmenter algorithm to find segment boundaries. This method was applied in TopicTiling algorithm [Riedl, 2012] and showed good results compared to other segmentation algorithms. The idea of this method is to calculate a similarity between left and right windows for each sentence ending. The sentence endings with the lowest values of this similarity are considered to be candidates for segment boundaries. Then some smoothing transformation is applied to the similarity function to obtain a so called *depth score*. The candidates with the depth score exceeding a certain threshold are selected as the final segment boundaries. The depth score can be also interpreted as a probability of the boundary in the corresponding sentence ending.

Our version of the segmenter algorithm differs from the original one, as we use sentence subjects to calculate similarities between the windows of sentences. In the original version, the authors used the topic IDs assigned to the words during the inference.

4.2. Segmentation quality metrics P_k and WindowDiff

P_k measure uses a sliding window with a length of k tokens, which is moved over the text to calculate the segmentation penalties. For each pair of words at a distance of k it is checked whether both words belong to the same segment or to different segments. This is done separately for the golden standard boundaries and the estimated segment boundaries. If the gold standard and the estimated segments do not match, a penalty of 1 is added. Finally, the error rate is computed by normalizing the penalty by the number of pairs. A value close to 0 denotes a perfect segmentation quality of the estimation.

The value of parameter k is assigned to half of the number of tokens in the document divided by the number of segments, given by the gold standard.

A drawback of the P_k measure is its unawareness of the number of segments between the pair of words. WindowDiff is an enhancement of P_k: the number of segments between the pair of words is counted. Then the number of segments is compared between the gold standard and the estimated segments. If the number of segments are not equal, 1 is added to the penalty, which is again normalized by the number of pairs to get an error rate between 0 and 1. [Riedl, 2012]

4.3. Artificially generated corpus description

We use PostScience corpus as a basis for the generated collection. We apply lemmatization, delete stop words and all documents that contain more than 200 sentences or less than 10 sentences. Then we compose artificial documents by concatenating full source documents. As it was mentioned in [Riedl, 2012], using full documents makes the corpus more realistic compared to the case when only the fragments of documents are concatenated.

To avoid topic repetition in sequential segments, we build a simple topic model on PostScience dataset and use only documents with different topics for sequential segments. Moreover, we use only the documents with the probability of one topic exceeding the threshold of 0.8. All of this allows us to assume that the golden standard segmentation boundaries also appear to be the topical boundaries.

The number of segments in a document varies from 2 to 4. The resulting number of documents in the generated corpus is 700.[1]

4.4. Experiments Setup

In all the provided experiments, BigARTM open-source library [Vorontsov, Frey, 2015] was used for topic model constructing. To find the optimal parameter values, we use the 5-fold cross-validation on the training subset. It includes 500 artificial documents. Here WindowDiff metrics is used for the evaluation.

Let us describe all the parameters that are the subject of our exploration:

- I is the number of iterations in EM-algorithm. This parameter is strongly connected with overfitting and model convergence.
- α is the strength of Theta sparsity regularizer. By tuning this parameter, we investigate whether a simple sparsity decreases the segmentation quality of topic models.
- τ_1 is the τ parameter for equation (3), which is used when segment boundaries match the sentence boundaries.
- τ_2 is the τ parameter for equation (3), which is used when the segment boundaries are being calculated iteratively by merging sequential segments.
- w is the size of window, which is used in segmenter algorithm to calculate final boundaries.

Table 1. Topic model parameter evaluation

Parameter	Optimal value	WindowDiff
I	40	0.253
α	0.2	0.248
τ_1	0.1	0.242
τ_2	11	0.232

[1] The artificially generated corpus ArtPostScience is available here: https://yadi.sk/d/fSswtwqV3SbsCD

The results of the parameters tuning are provided in Table 1. Once the the optimal number of iterations was found, we define the structure of the training process. During the first 5 iterations the topic model works without any regularizers. At the 5th iteration the regularizer of Θ sparsity starts to work. For the last 25 iterations, we apply equation (3) on each E-step of the algorithm. This structure of training is essential, because the ability to use segment subjects requires the convergence of the topic model.

As we can see from Table 1, the optimal value of α parameter is very small. That means that strong influence of Θ sparsity regularizer lowers the ability of the topic model to restore the segment boundaries.

One can also note, that iterative merging of segments gets a high τ_2 coefficient in equation (3), while the strategy with fixed sentence boundaries keeps this coefficient low. That means that making sparse small segments like sentences is undesirable. That can be explained by the fact that small sentences depend more on separate words topics and their subjects are more unstable.

The optimal value for the window parameter was 11. It may be explained by the fact that the shortest gold standard segments in our collection have the size of 10 sentences. Thus, in real collection we would recommend to set this parameter equal to the length of the shortest segment.

5. Main results

All the models with the optimal parameter values were evaluated on the test documents of our artificial dataset. The number of test documents is 200. The train documents were used only to build the topic model.

The final results in segmentation are shown in table 2 and compared against TopicTiling model from [Riedl, 2012]. For TopicTiling baseline model, we reproduce the original estimation of similarities between windows based on topic assignments in LDA inference. In our models, we use sentence subjects to calculate the similarities. *PLSA + Θ sparsity* is the model that uses Θ sparsity regularizer. *PLSA + SentenceSparse* is the model that applies equation (3) using sentence subjects. *PLSA + SegmentSparse* is the model with iterative segment merging. We used estimated parameters for all models provided.

Table 2. Final results in segmentation

Model	WindowDiff metrics	P_k metrics
TopicTiling	0.258	0.145
PLSA + Θ sparsity	0.173	0.100
PLSA + SentenceSparse	0.159	0.099
PLSA + SegmentSparse	**0.155**	**0.095**

One can see that the proposed method with regularization (3) and iterative merging of segments gives the best segmentation quality. Also both TopicTiling and PLSA + Θ-sparse models, that are built in a bag-of-words assumption, show the worst results in segmentation.

To better explore the effective number of topics in the models, we compare the sparsity levels of Θ matrix in Table 3. The SegmentSparse model decreases the average number of topics almost in 3 times. Without implementation of equation (3), such result in sparsity could have only been achieved with a loss in segmentation quality.

Table 3. Results of Θ sparsity for different models

Model name	Ration of non-zeros in Θ
PLSA + Θ sparsity reg.	4%
PLSA + SentenceSparse	1.8%
PLSA + SegmentSparse	**1.5%**

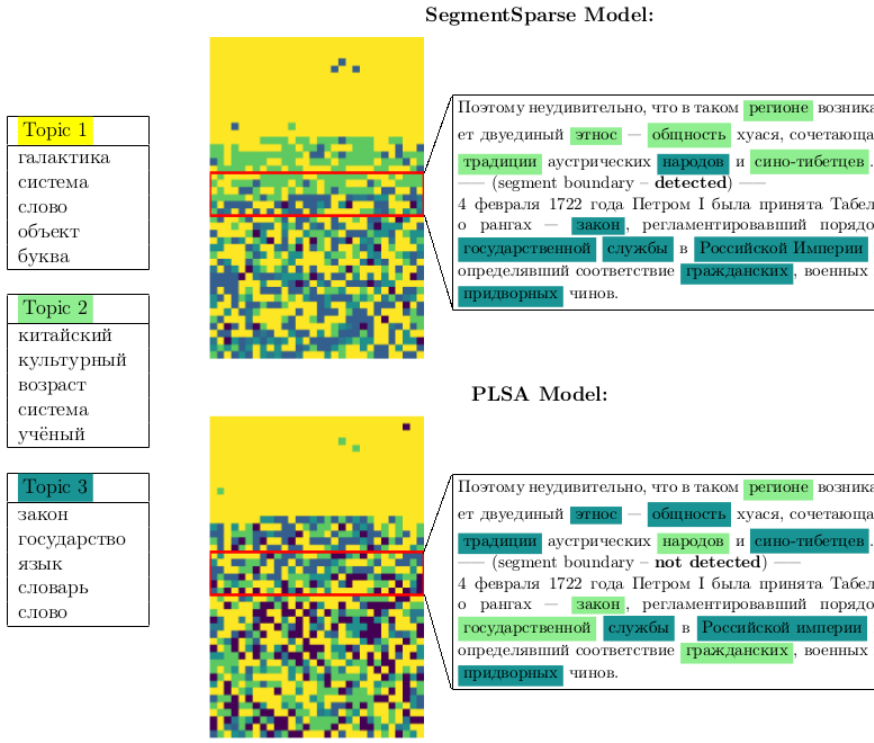


Figure 1. The visualization of PLSA + Θ sparsity reg. (bottom) and SegmentSparse (top) models applied to the same test document. Words on the figure are represented with pixels, which follow each other from left to right and from top to bottom. The text fragment where the PLSA model has failed is marked with red.

Let us look more closely into how topics are spread along a document. In Fig. 1 we represent dominant topics for each word in the sequential text with a color. The PLSA + Θ-sparse model reveals only two semantic segments in the text

and in the second segment all the topics are mixed up. Whereas the SegmentSparse model catches all three segments and makes them topically different. Yellow topic appears in both second and third segments so we don't mark its words in the given text fragments. Also note, that the second model used up 32 topics, while for the first model 9 topics were enough. So in this document the SegmentSparse model outperformed PLSA + θ-sparse in both sparsity and segment boundaries estimation.

Now we provide the results of the SegmentSparse model application to the documents of the original PostScience collection. We still use the model that was trained on artificial documents. On Fig. 3 we represent an original document, where SegmentSparse model found two different segments. The first topical segment contains historical information when the second segment is more about natural features of the region. So this segmentation of the document is justified.

... Казанская губерния, наоборот, вошла по просьбе Коржинского. Это интересная историческая ситуация, но постепенно, уже к 20-м годам XX века средняя Россия охватывает территорию от Ярославской и Костромской губерний на севере до Воронежской и Саратовской на юге. Вот эта вся территория находится в пределах европейской части России на левобережье Волги. В среднем, по некоторым подсчётам, природная флора этого региона насчитывает примерно 4,5 тысячи видов, очень немного. Для сравнения флора Турции, которая меньше по площади, включает больше 15 тысяч видов. Связана бедность флоры, с одной стороны, с тем, что это равнинная территория...

Figure 3. The topic representation of an original document from PostScience where SegmentSparse model found 2 segments

6. Conclusion

In this work we have shown, that going beyond the bag-of-words assumption in topic modeling gives a significant improvement in determining text structure. Our enhancements of the EM-algorithm allow to consider words co-occurrences without complicated modifications of the iterative process. Furthermore, the proposed method increases sparsity of documents subjects. This means, the topic models we have built are simpler and provide a better and more interpretable text representation than the models that are built within the bag-of-words assumption. Our iterative segments merging procedure highly increases the segmentation quality of the topic model. As it was shown, the model's confidence in segment boundary identification increased.

For the further research, we are going to implement more assumptions about topic structure in texts, following the same approach. Besides, we would like to focus on the applications of such topic models. Better reflection of the real text structure in a topic model can bring significant improvements to many down-stream tasks.

7. Acknowledgements

The work was supported by Government of the Russian Federation (agreement 05.Y09.21.0018) and the Russian Foundation for Basic Research grant 17-07-01536.

References

1. *Blei D. M., Ng A. Y., Jordan M. I.* (2003), Latent Dirichlet allocation, Journal of Machine Learning Research, Vol. 3, pp. 993–1022.
2. *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.
3. *Galley M., McKeown K., Fosler-Lussier E., Jing H.* (2003), Discourse Segmentation of Multi-Party Conversation, ACL '03 Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Vol. 1, pp 562–569
4. *Vorontsov K., Potapenko A.,* (2014). Additive regularization of topic models. Machine Learning. 101. 1–21. 10.1007/s10994-014-5476-6.
5. *Vorontsov K. V.* (2014), Additive Regularization for Topic Models of Text Collections, Doklady Akademii Nauk, Vol. 455, no. 3.
6. *Vorontsov K., Frey A., Romov P., Yanina A., Suvorova M., Apishev M.* (2015) Bigartm: open-source library for topic modeling of big text collections [Bigartm: biblioteka s otkritim kodom dlya tematicheskogo modelirovaniya bolshich tekstovich kollektsiy] In Analytics and data management in areas with intensive use of data [Analitika I upravleniye v oblastyach s intensivnim ispolzovaniem dannich]. DAMDID/RCDL'2015, Obninsk, pp 28–36.
7. *Martin Riedl, Chris Biemann.* (2012), Text Segmentation with Topic Models. Journal for Language Technology and Computational Linguistics (JLCL) Vol. 27 — pp. 47–69
8. *Lan Du, Wray Buntine, Mark Johnson,* (2013), Topic Segmentation with a Structured Topic Model, Proceedings of NAACL-HLT 2013, pp. 190–200
9. *J. Pitman and M. Yor.* (1997) The two-parameter Poisson-Diriclet distribution derived from a stable subordinator. Annals Probability, Vol.25, pp. 855–900.
10. *G. Balikas, M. Reza Amini, M. Clausel,* (2016). On a Topic Model for Sentences. 10.1145/2911451.2914714.

BUILDING A CORPUS FOR THE QUANTITATIVE RESEARCH OF RUSSIAN DRAMA: COMPOSITION, STRUCTURE, CASE STUDIES

Skorinkin D. (dskorinkin@hse.ru),
Fischer F. (frafis@gmail.com),
Palchikov G. (rebel368@gmail.com)
National Research University Higher School of Economics

In this paper we introduce RusDraCor—an open corpus of Russian drama for digital literary & linguistic research. The corpus (rus.dracor.org) contains plays from the middle of XVIII to the first third of XX century provided with structural (plus some semantic) markup and metadata. Texts are encoded in the XML-based standard TEI, widely used in building corpora for the humanities. We describe the contents and annotation layers of our corpus, provide some details on its development and enrichment, and finally describe three research cases. Each case demonstrates the use of RusDraCor to answer specific questions about composition, structural features and historical evolution of Russian drama.

Keywords: corpora, TEI, XML, markup, drama, Russian drama, digital humanities, digital literary studies, stylometry, network analysis

РАЗРАБОТКА КОРПУСА ДЛЯ АНАЛИЗА РУССКИХ ДРАМАТИЧЕСКИХ ТЕКСТОВ: СОСТАВ, СТРУКТУРА, ИССЛЕДОВАТЕЛЬСКИЕ СЦЕНАРИИ

1. Introduction

The development of richly-encoded dramatic corpora for digital literary research has been on the rise recently. Some examples include “Shakespeare His Contemporaries” (510 English plays from the Shakespeare era, [Mueller 2014]) “Théâtre Classique” (1080 French dramas from the XVII and XVIII century, collected and encoded by Paul Fièvre at http://www.theatre-classique.fr/), the DLINA corpus (466 German-language plays from 1730 up to 1930, [Fischer et al. 2016a]) , the Dramawebben

(66 Swedish plays at http://dramawebben.se). These corpora are encoded in TEI and have all proved their usefulness for the digital literary studies [Glorieux 2016], [Xanthos et al. 2016], [Fischer et al. 2016a,b]. Adding a Russian-language collection to the family of drama corpora will enable similar research on Russian material and boost cross-cultural studies on the structure and evolution of dramatic texts.

The ultimate goal of the RusDraCor project is to provide a corpus of at least 500 encoded Russian plays spanning for two centuries, roughly in between 1740 and 1940 (later plays are still under copyright). Currently the corpus (rus.dracor.org) features 89 plays, provided with semantic and structural annotation described below. The earliest play in RusDraCor is Horev (Хорев) by A. P. Sumarokov (1747); the newest—Ivan Vasilievich (Иван Васильевич) by M. A. Bulgakov (1936). The main sources for growing the corpus are Wikisource (wikisource.org), the Russian Virtual Library (rvb.ru), Online library of Alexei Komarov (ilibrary.ru) and Maxim Moshkov's library (lib.ru).

2. Annotation and metadata available at RusDraCor

RusDraCor provides both structural and semantic markup for the plays included. It also contains certain meta information about the encoded texts. The corpus is encoded in accordance with the TEI guidelines (http://www.tei-c.org/Guidelines/), a widely used 30-year old XML standard comprising around 550 elements, specifically defined for digital editions and the demands of Digital Humanities research. Source TEI/XML files of the corpus are available at https://github.com/dracor-org/rusdracor. In this section we describe the layers of the corpus annotation implemented at this moment.

2.1. Structural markup

The structural markup asumes the hierarchical representation of all subdivisions in the play (acts, scenes, enters etc.). This is done with help of div tag:

```
<div type="act">
<head>Действие первое</head>
<div type="scene">
<head>Сцена первая</head>
<div type="enter">
<head>Выход первый</head>
<stage>Игроки, князь Звездич, Казарин и Шприх. За столом мечут банк
        и понтируют...
Кругом стоят.</stage>
<!--..text of the first enter..-->
</div>
</div>
</div>
```

The main components of a dramatic texts are character speeches and stage directions. Our markup uses TEI tags <p> (paragraph) or <lg> + <l>(verse line group and a single verse line) for speeches and TEI tag <stage> for stage directions. Speaker

is encoded with <speaker> tag; each character gets a unique identifier by which s/he is referenced in the markup throughout the entire play:

```
<stage>Городничий, попечитель богоугодных заведений, смотритель
        училищ, судья, частный пристав, лекарь, два квартальных.
        </stage>
<sp who="#Gorodnichij">
<speaker>Городничий.</speaker>
<p>Я пригласил вас, господа, с тем чтобы сообщить вам пренеприятное
    известие: к нам едет ревизор.</p>
</sp>
<sp who="#AmmosFedorovichLjapkinTjapkin">
<speaker>Аммос Федорович.</speaker>
<p>Как ревизор?</p>
</sp>

***

<sp who="#Famusov">
<speaker>Фамусов</speaker>
<lg>
<l>Сказал бы я, во-первых: не блажи,</l>
<l>Именьем, брат, не упрекай оплошно,</l>
<l>А, главное, поди-тка послужи.</l>
</lg>
</sp>
<sp who="#Chatskij">
<speaker>Чацкий</speaker>
<lg>
<l>Служить бы рад, прислуживаться тошно.</l>
</lg>
</sp>
```

Cases of multiple speech authorship ('Все', 'Оба', 'Вместе') are resolved manually if possible at all:

```
<sp who="#Zagoretskij #PervajajaKnjazhna #VtorajajaKnjazhna #TretjajaKnjazhna #ChetviortajajaKnjazhna #PjatajajaKnjazhna #ShestajajaKnjazhna">
<speaker>Все вместе</speaker>
<lg>
<l>Мсьё Репетилов! Вы! Мсье Репетилов, что вы!</l>
<l>Да как вы! Можно ль против всех!</l>
<l>Да почему вы? стыд и смех.</l>
</lg>
</sp>
```

2.2. Semantic markup

As of now, semantic markup is mostly limited to the specification of each character's gender. Gender is first assigned automatically to each character during the initial conversion (relying on typical name endings). Then it goes through manual correction. We use standard way of specifying gender via the @sex attribute of each <person> element in <listPerson> of the <teiHeader> that is recommended by the TEI Consortium (see the 'TEI Header' section of the TEI P5: Guidelines at http://www.tei-c.org):

```
<listPerson>
<person xml:id="MarijaVasilevna" sex="FEMALE">
<persName>Мария Васильевна</persName>
<persName xml:lang="en">Mariâ Vasil'evna</persName>
<persName xml:lang="de">Mariâ Vasil'evna</persName>
</person>
<person xml:id="Telegin" sex="MALE">
<persName>Телегин</persName>
<persName xml:lang="en">Telegin</persName>
<persName xml:lang="de">Telegin</persName>
</person>
<!-- … rest of the list .. -->
<listPerson>
```

We are also working on adding the 'social status' information for each character—whether s/he belongs to the nobility, or is a servant, a serf, a soldier, a merchant and so on. This could give more material for formal analyses of character systems that we implement (see the research cases below).

2.3. Metadata

The metadata is stored in the <teiHeader> element of each document. It contains information about the author; dates of origin, publication and premiere of the play (if available), character names and IDs, link to the source of the text. The following example demonstrates a part of a play metadata containing the information about the play and its author:

```
<titleStmt>
<title type="main" xml:lang="ru">Гроза</title>
<title type="main" xml:lang="en">The Storm</title>
<title type="sub" xml:lang="ru">Драма в пяти действиях</title>
<title type="sub" xml:lang="en">A Drama in Five Acts</title>
<author key="Wikidata:Q171976">Островский, Александр Николаевич
</author>
</titleStmt>
```

Second example shows the encoding of metadata related to dates of creation, premiere and first publication:

<bibl type="originalSource">
<title>А. Грибоедов. Горе от ума. А. Сухово-Кобылин. Пьесы. А. Островский. Пьесы. "Библиотека Всемирной литературы", М.: Художественная литература, 1974.</title>
<date type="print" when="1860">1860 год (wikipedia)</date>
<date type="premiere" when="1859">1859 год (wikipedia)</date>
<date type="written" when="1859">1859 год (wikipedia)</date>
</bibl>

3. Case study 1. Measuring the evolution of drama through stage directions

3.1. Rationale for the research

Our first case study is a research of the evolution of dramatic texts. We demonstrate the use of RusDraCor, which currently contains plays written in between the middle of XVIII and the first third of the XX centuries, for diachronic studies. Specifically, we analyze the changes in length and linguistic composition of *stage directions* (<stage> tag, see the 'Structural markup' section above). These changes, in our view, reflect the general 'epification' of drama—a process that later reaches its peak with the emergence of Brecht's 'epic theatre' theory [Брехт 1965].

"A stage direction can be detailed and evocative <…> More typically, however, is direction that lacks specific details but instead invokes a formula where the implementation of the onstage effect is left to the players or to the imagination of a reader" [Dessen 2011]. When one reads a play from the XVIII of early XIX century, s/he may not even notice stage directions at all. They are typically short and purely technical:

1) <stage>Те же и невольник.</stage> (А. П. Сумароков. Хорев. 1747)
2) <stage>Тамира и Клеона.</stage> (М. В. Ломоносов. Тамира и Селим. 1750)
3) <stage>Оскольд, Семира, Избрана, Возвед и воины.</stage> (А. П. Сумароков. Семира. 1751)
4) <stage>Слуги уходят.</stage> (А. А. Шаховской. Пустодомы. 1819)
5) <stage>Народ расходится.</stage> (А. С. Пушкин. Борис Годунов. 1826)

However, as new types of drama evolve, stage directions become more elaborate and content-rich, turning into a significant part of the dramatic narrative. Consider these examples from plays in our corpus:

6) <stage>Слышно, как к дому подъезжают два экипажа. Лопахин и Дуняша быстро уходят. Сцена пуста. В соседних комнатах начинается шум. Через сцену, опираясь на палочку, торопливо проходит Фирс, ездивший встречать Любовь Андреевну; он в старинной ливрее и в высокой шляпе; что-то говорит сам с собой, но нельзя разобрать ни одного слова. Шум за сценой все усиливается. Голос: «Вот пройдемте здесь...»

Любовь Андреевна, Аня и Шарлотта Ивановна с собачкой на цепочке, одетые по-дорожному. Варя в пальто и платке, Гаев, Симеонов-Пищик, Лопахин, Дуняша с узлом и зонтиком, прислуга с вещами—все идут через комнату.</**stage**> (А. П. Чехов. Вишневый сад. 1905)

7) <**stage**>Прыгает в окно. Даль, видимая в окне, оказывается нарисованной на бумаге. Бумага лопнула. Арлекин полетел вверх ногами в пустоту. В бумажном разрыве видно одно светлеющее небо. Ночь истекает, копошится утро. На фоне занимающейся зари стоит, чуть колеблемая дорассветным ветром,—Смерть, в длинных белых пеленах, с матовым женственным лицом и с косой на плече. Лезвее серебрится, как опрокинутый месяц, умирающий утром. Все бросились в ужасе в разные стороны. Рыцарь споткнулся на деревянный меч. Дамы разроняли цветы по всей сцене. Маски, неподвижно прижавшиеся, как бы распятые у стен, кажутся куклами из этнографического музея. Любовницы спрятали лица в плащи любовников. Профиль голубой маски тонко вырезывается на утреннем небе. У ног ее испуганная, коленопреклоненная розовая маска прижалась к его руке губами. Как из земли выросший Пьеро медленно идет через всю сцену, простирая руки к Смерти. По мере его приближения черты Ее начинают оживать. Румянец заиграл на матовости щек. Серебряная коса теряется в стелющемся утреннем тумане. На фоне зари, в нише окна, стоит с тихой улыбкой на спокойном лице красивая девушка—Коломбина. В ту минуту, как Пьеро подходит и хочет коснуться ее руки своей рукой, между ним и Коломбиной просовывается торжествующая голова автора.</**stage**> (А. А. Блок. Балаганчик. 1906)

8) <**stage**>Грохот, взрыв, выстрел. Победоносиков распахивает дверь и бросается в квартиру. На нижней площадке фейерверочный огонь. На месте поставленного аппарата светящаяся женщина со свитком в светящихся буквах. Горит слово «Мандат». Общее остолбенение. Выскакивает Оптимистенко, на ходу подтягивает брюки, в ночных туфлях на босы ноги, вооружен. </**stage**> (В. В. Маяковский. Баня. 1929)

In a manner similar to the study of the evolution of novelistic titles by Moretti [Moretti 2009], we made an attempt to quantify and measure these changes in dramatic texts.

3.2. Analysis

To measure the evolution of stage directions in plays over time, one could use relatively simple & obvious features. Given the examples above, one obvious choice would be to use a set of features measuring absolute and relative lengths of the stage directions. We implemented the following measures (calculated for *each play*):

1. total length of stage directions in a play (figure 1)

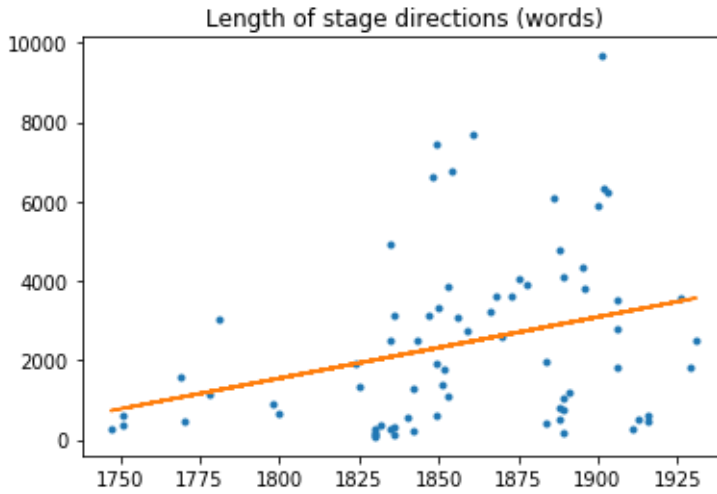


Figure 1

2. average length of a stage direction (figure 2)

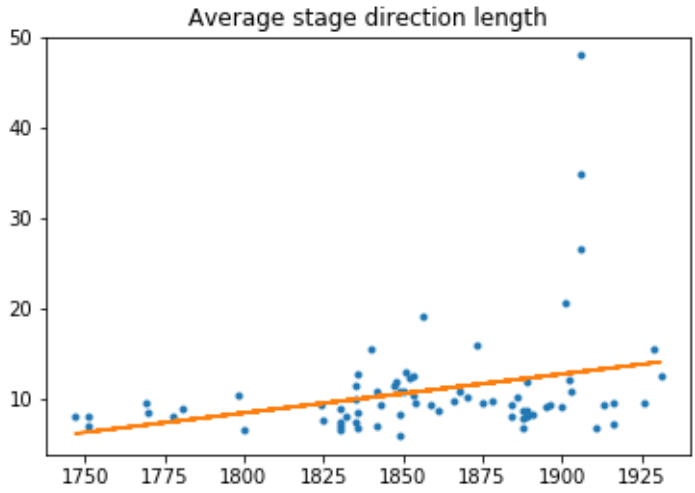


Figure 2

3. ratio of stage directions text to the direct speeches text (measured in word tokens). (figure 3)

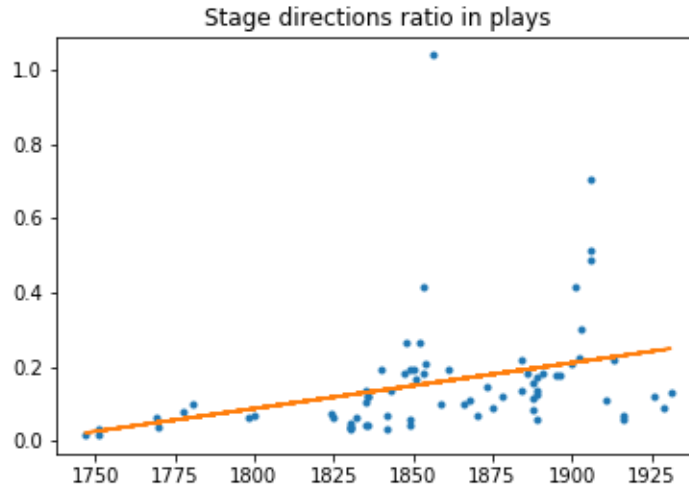


Figure 3

Another set of features verb usage in stage directions (we used MyStem (https://tech.yandex.ru/mystem/) to obtain PoS tags). As one may notice, earlier stage directions (examples 1–5) seem to contain only few verbs. These verbs usually describe the technical dynamics of the play: characters *entering* or *leaving*, also in some cases *laughing*, *crying*, *dying* and so on. In later stage directions (examples 5–8) there is a greater abundance diversity of verbs, which can be considered a marker of a *narrative* stage direction. Therefore, we also implemented the following verb-related measures:

1. the total share of verbs per all words in stage directions texts in a play (figure 4)

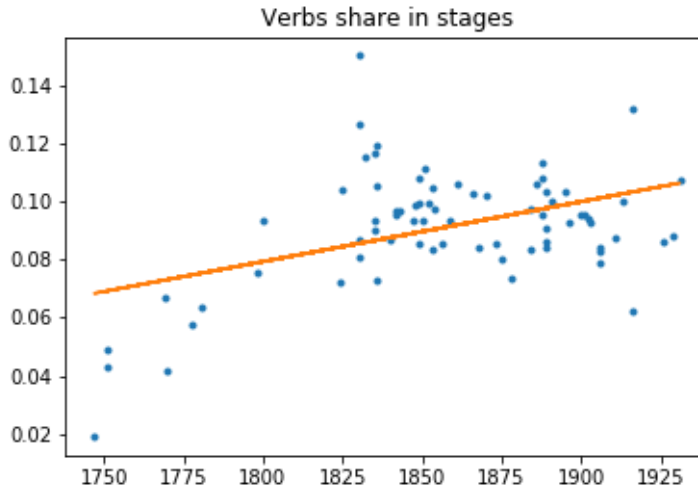


Figure 4

2. the total number of unique verbs in all stage directions (figure 5)

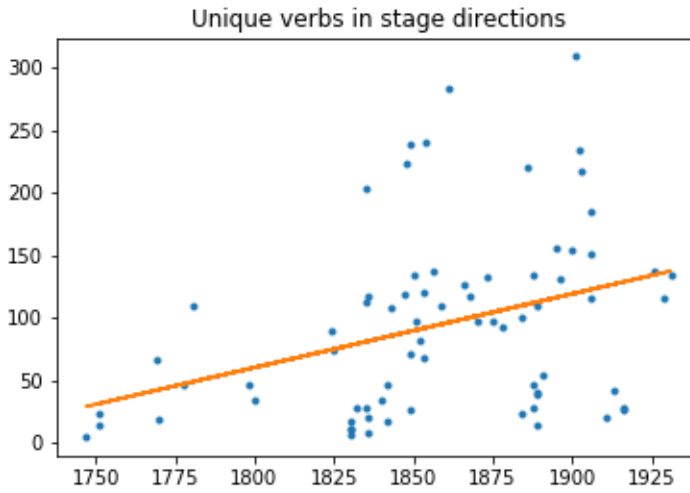


Figure 5

As one can see from figures 1–5, all measures show steady increase over time. And though the dependency is not linear in the strict sense, we can at least claim that no play in XVIII or early XIX century has long diverse stage directions of the narrative kind. In the late XIX and early XX centuries, on the other hand, we have a lot of plays with traits of 'epification' in them. Of course, this is only preliminary research, but the result could be a visible trace of a cultural evolutionary process that lead to the emergence of Brecht's epic theatre theory.

4. Case study 2. Gender specifics in character speech

4.1. Rationale for the research

In our second case study we switch to the analysis of direct speech in drama. Using the structural markup for speeches (<sp> tag and @who attribute) we can easily extract every speech instance for each character. And since we have gender information in the metadata, one obvious research goal could be to perform the statistical comparison of male and female speeches in the corpus. Similar research on movie subtitles [Schofield, Mehr 2016] produced fruitful results.

4.2. Tools and preprocessing

The earliest quantitative work on character speech is probably [Burrows 1987]—the now-famous book that laid the foundations of contemporary stylometry

(computational stylistics). In this research we also chose to use stylometric tools, namely the widely used Stylo package for R [Eder et al. 2016]. Stylo has a number of built-in stylometric functions for statistical exploratory analysis of differences in text/speech styles. To eliminate the effect of morphology we performed analysis on lemmatized text, using MyStem (https://tech.yandex.ru/mystem/) for lemmatization.

4.3. Analysis

To perform contrastive analysis of male and female speech in our corpus we used the oppose () function of Stylo package. This function performs a contrastive analysis between two given sets of texts, using Burrows's Zeta [Burrows 2007] and its extensions by [Craig & Kinney, 2009]. The function takes two sets of texts as input and outputs words significantly preferred and avoided by texts in one set (as compared to the other). Figure 6 shows such words for female speech (most preferred words are top left, most avoided words, as compared to male speech,—bottom right).

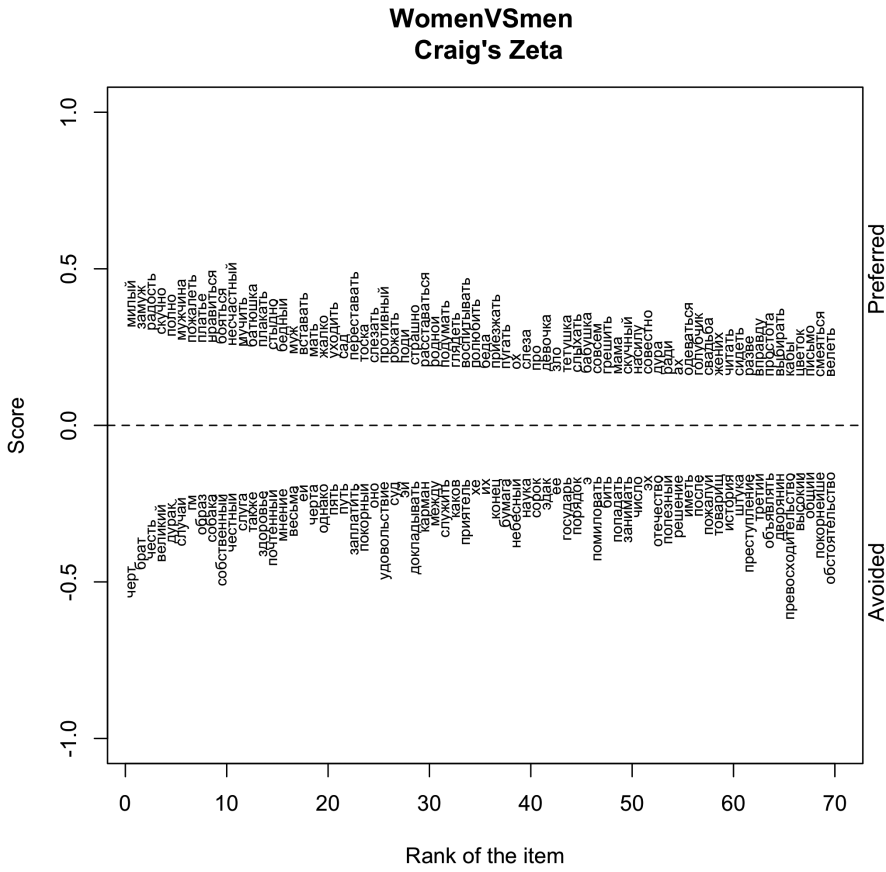


Figure 6. Results of the oppose() function applied to male and female character speech in RusDraCor

As one can see in figure 6, statistical analysis demonstrates results that can be called (awfully) stereotypical. Women tend to talk about marriages, matrimonial activity and procreation (‘мужчина’ = man, ‘муж’=husband, жених = ‘groom’, ‘замуж’= (to get) married, ‘рожать’=to give birth, ‘дитя’=child, ‘воспитывать’= to bring up (a child)), feelings & emotions (‘радость’ = joy, ‘весело’=cheerfully, ‘стыдно’=ashamed, ‘жаль’= it’s a pity, ‘счастие’ =happiness), clothes (‘платье’=dress, ‘туалет’=clothes, ‘одеваться’= get dressed) family (‘бабушка’=grandmother, ‘мать’=mother, папа=daddy, ‘маменька’=mummy, ‘тетушка’=auntie). Men, meanwhile, use swear words and offensive language (‘черт’=devil/demon, usually an expressive interjection, ‘дурак’=fool), talk about honor, honesty, affairs of the state and government service (‘бумага’= paper/official document, ‘превосходительство’= (your) excellency, standard address to an official of a certain rank, ‘государь’=emperor/polite address to a person, ‘докладывать’=to make an official report to a superior, ‘служить’=to serve, ‘служба’=service).

5. Case study 3. Network analysis

5.1. Rationale for the research

Literary network analysis is a sub-branch of digital literary studies that applies methods of network science to the study of fiction. The rise of literary network analysis is typically associated with the works of Moretti, who provided the philological rationale for this sort of digital formalism in [Moretti 2011] using Shakespeare’s *Hamlet* as a showcase. However, there is also substantial amount of earlier research dedicated to network analysis of literary work. In [Schweize & Schnegg 1998] anthropologists analyze the network of characters in *Simple stories*, a contemporary novel by Ingo Shulz describing life in the former GDR after the unification of Germany. [Alberich et al 2002] explore the vast network of Marvel comics characters, extracted automatically from a total of 12,942 comics issues. The authors apply theoretical apparatus from graph theory, namely network density, clustering coefficients, average node degree, average path length and other formal metrics of the resulting network. This study demonstrates that fictional networks are structurally similar to the social networks of the real world and can be investigated with help of standard approaches from social network analysis.

In a follow-up study of the same Marvel universe [Gleiser 2007] all characters are additionally classified into heroes and villains, which enables authors to speculate on Marvel’s marketing techniques. [Gleiser 2007] demonstrate that most heroes are connected to each other within one huge connected component of the network, whereas villains do not form a unified group. This, the authors suggest, could result from Marvel’s attempts to popularize new and yet unknown characters by pairing them with older well-known superheroes, such as Captain America or Superman.

Other early network-related research includes several analyses of Shakespeare’s plays [Stiller et al. 2003] [Stiller & Hudson 2005], analysis of community

structures in *Les Miserables* [Newman & Girvan 2004], comparison of rural and urban networks in XIX century British novels [Elson et. al 2010]. After [Moretti 2011] a lot more research on literary network analysis came around, see [Agarwal et al. 2012], [Lee & Yeung 2012], [Agarwal et al. 2013], [Ardunay & Sporleder 2014], [Lee & Wong 2016], [Grayson et al. 2016]. Literary network studies on Russian material include [Bodrova & Bocharov 2014] and [Skorinkin 2017].

Dramatic text with its inherent structure (acts, scenes, speeches) naturally becomes an easier target for automated network extraction and analysis. Studies like [Trilcke et al. 2015], [Trilcke et al. 2016], [Fischer et al. 2017] employ network analysis to large-scale digital exploration of drama (in a way following Moretti's lead with *Hamlet*).

5.2. Extracting networks

In our research we follow older formalist/structuralist approaches in literary studies [Ярхо 1997 (1930-ies)], [Сапогов 1974], [Лотман 1998]. We formalize interactions in drama as co-appearance of two characters in one scene of a play, in which both character speak at least once. This formalisation has its drawbacks, but its huge benefit is that it allows easy conversion of a play (provided with structural markup) into a network of characters and their interactions. And the availability of a multitude of plays in our corpus opens opportunities for large-scale research on the structure and evolution of different compositional types of plays. Figure 7 (also available attached as a separate file in scalable SVG format) contains a visualization of character networks extracted from all the plays currently included in RusDraCor.

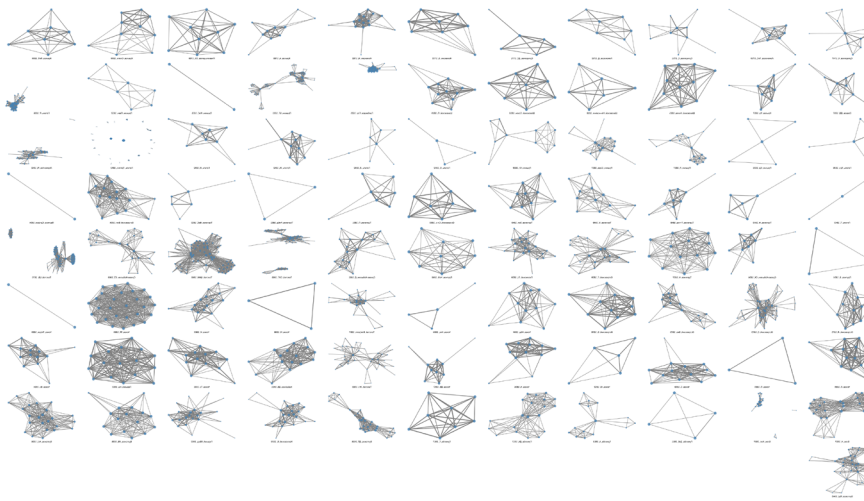


Figure 7. Visualization of character networks extracted from all plays currently included in RusDraCor (ordered chronologically)

5.3. Large-scale network analysis

Even simple visual analysis of figure 7 already tells something about certain changes in types of drama over time. For instance, one may notice from the first two rows that the networks of plays from XVIII and early XIX century all share certain traits. These traits include

1. relatively small number of nodes (characters)
2. single densely interconnected core, with few to none periphery characters (which are typically servants)

This structure apparently reflects the classicist tradition with its three unities of action, time and place. For better demonstration we provide figure 8, which features all plays in our corpus from 1747 (earliest) to 1825.

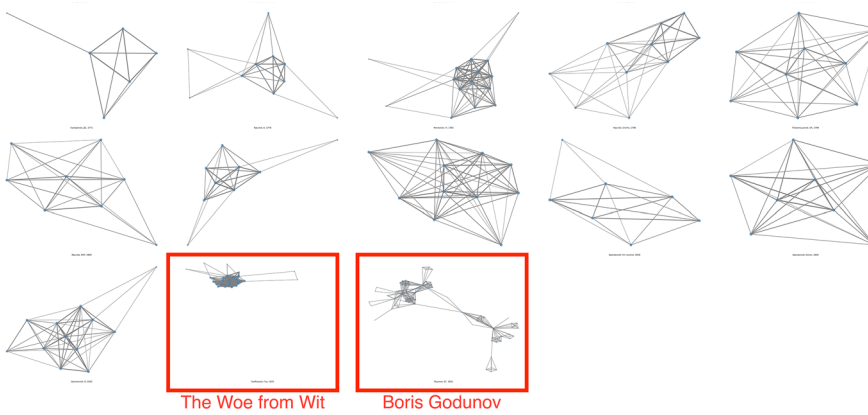


Figure 8. Visualization of character networks for all plays in our corpus from 1747 (earliest) to 1825. Ordered chronologically

The first two plays in our corpus to violate the standard structure are (marked with red) Griboedov's Woe from Wit ("Горе от ума") and especially Pushkin's Boris Godunov (Борис Годунов). Both plays are known to be a result of Shakespearean influence, and Shakespeare himself was an acknowledged breaker of the classical tradition [Dryden 1668]. Similar observations were made in [Fischer et al. 2016b] with regard to Shakespearean influence on Goethe and the structural evolution of German drama. A huge advantage of network formalization is the possibility to combine visual analysis with strict mathematical measures provided by graph theory. The visible difference of mentioned networks can be observed through networks metrics. Some of the metrics are

1. number of nodes
2. network density, which is the ratio of the number of edges in a graph to the maximum possible number of edges in that graph (i.e. if each node was connected to every other node).
3. network diameter, or the length of the longest path between one node and another in that network, measured in the number of edges

Figures 9, 10 and 11 present number of node, density and diameter measures for each play in our corpus. NB that networks with several components have no diameter measure in this implementation.

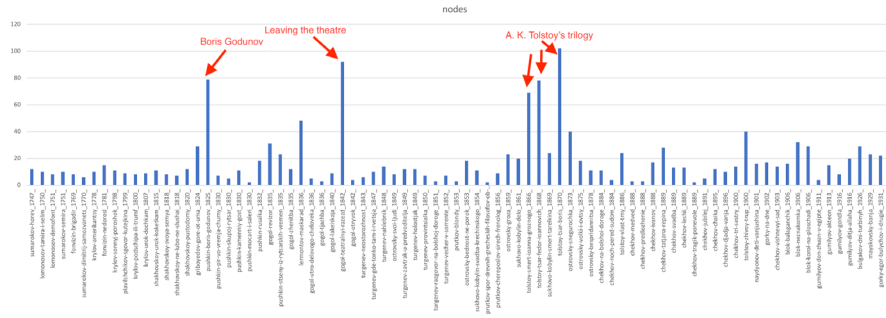


Figure 9. Number of nodes in play network

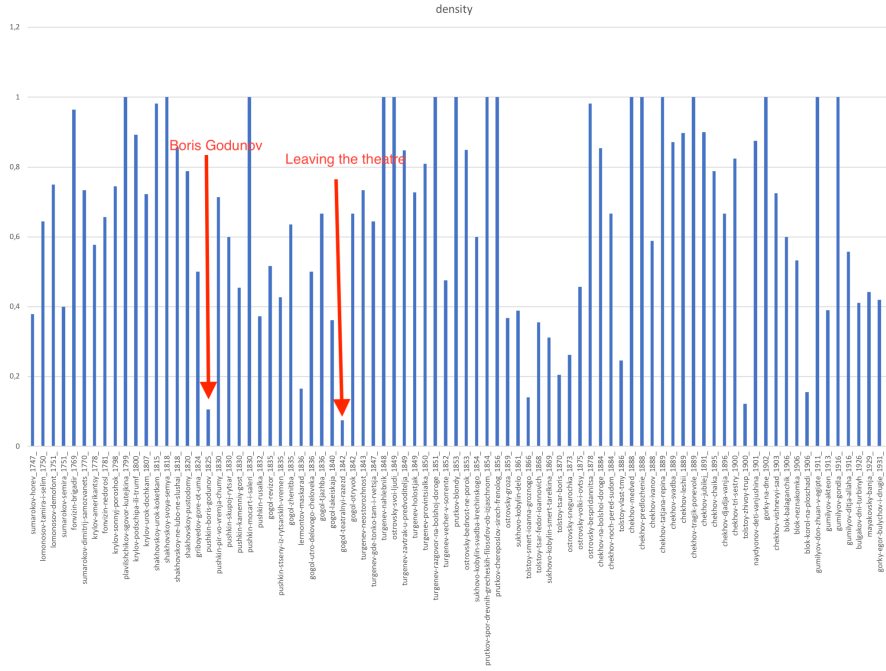


Figure 10. Network densities

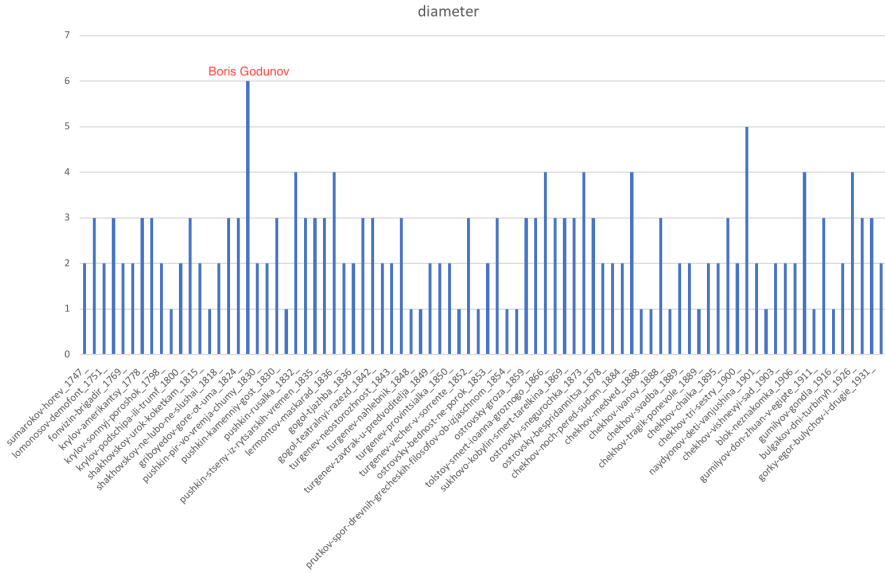


Figure 11. Network diameters

As one can see, the networks that are visually different also have extreme network measures. For instance, Boris Godunov, chronologically the first play with no single 'core' of main characters (see figure 7), is clearly an outlier in terms of the number of characters (much higher than the others), density (much lower) and diameter (much bigger). All these measures obviously reflect the specific structure of Pushkin's play, the fact that its plot takes place in different chronotopes and that the play itself was not meant to be staged (a specimen of the 'closet' plays). It is also no accident that similar measures are observed in the dramatic trilogy by A. K. Tolstoy, which also exhibits Shakespearean traits. Another example of a play with many characters and low density is Gogol's Leaving the theatre (Театральный разъезд)—a very specific meta-play of peculiar structure.

Other network measures can be used to track the evolution of plays in general. For instance, on figure 9 one can see that the average degree of a node in play (that is the number of connections each character has according to the chosen formalisation) gradually increases over time. This could also signify important changes in the types of drama produced in different time periods.

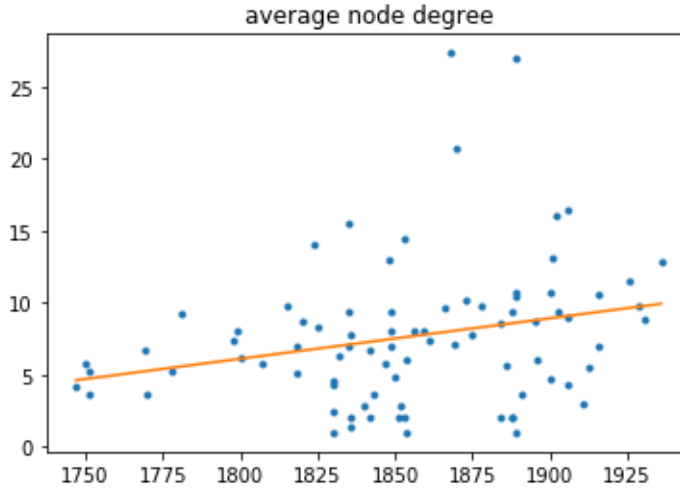


Figure 12. Average node degree in plays

5.4. Zooming in on Boris Godunov

All the research described above represents the Distant Reading approach [Moretti 2013] to literary studies, where distance (i.e. large-scale quantitative analysis) "is a condition of knowledge" [Moretti 2013: 129]. However, some people advocate a less radical, blended approach called Scalable Reading [Weitin 2017], where after large-scale analysis researcher might zoom in onto interesting samples. Here we take this step with Boris Godunov.

As we demonstrated above, Pushkin's play is one of the most expressive outliers in our corpus. The structure of the network, with several clearly distinguishable clusters, makes it an interesting target for more detailed network analysis.

Figure 13 shows the network for Boris Godunov visualized with Gephi [Bastian et al. 2009]. Colors represent automatic modularity clustering, which obviously captures Polish cluster with the False Dimitry in the middle, Moscow cluster with tsar Boris and his son Feodor, and the People cluster ('People/Народ' is an important 'group' character in the play). Node sizes are proportional to weighted degree, and the most central nodes generally correspond to the main characters (False Dimitry, Feodor, Boris etc.).

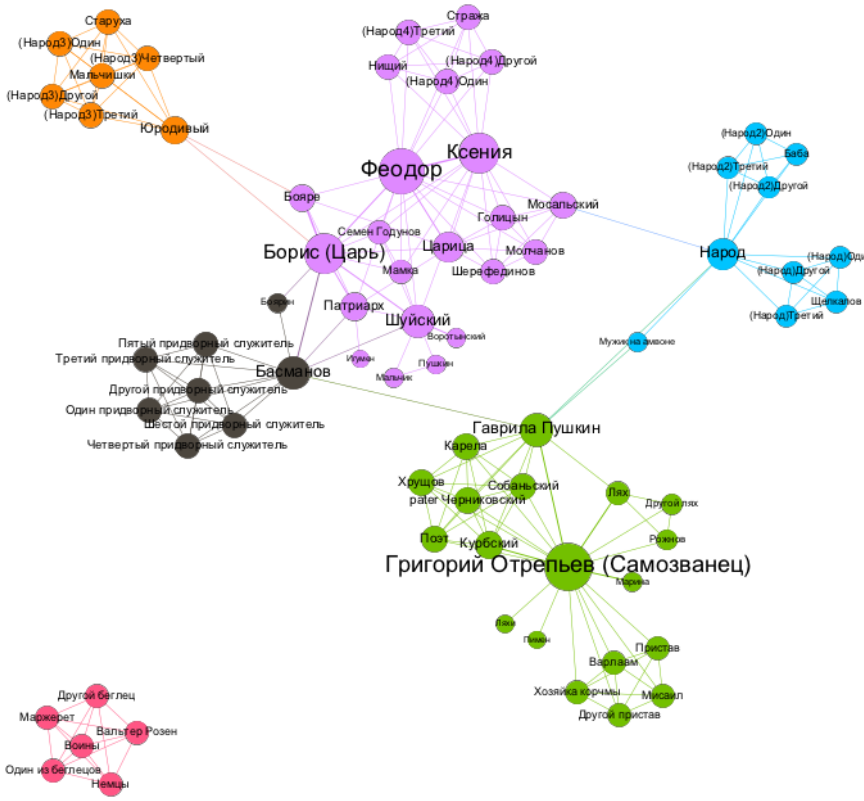


Figure 13. Boris Godunov network, nodes proportional to weighted degree

However, if we change the preferred centrality measure to betweenness centrality (нагрузка узла), i.e. the number of shortest paths going through this node, the picture changes significantly. The alternative visualization can be seen on figure 14.

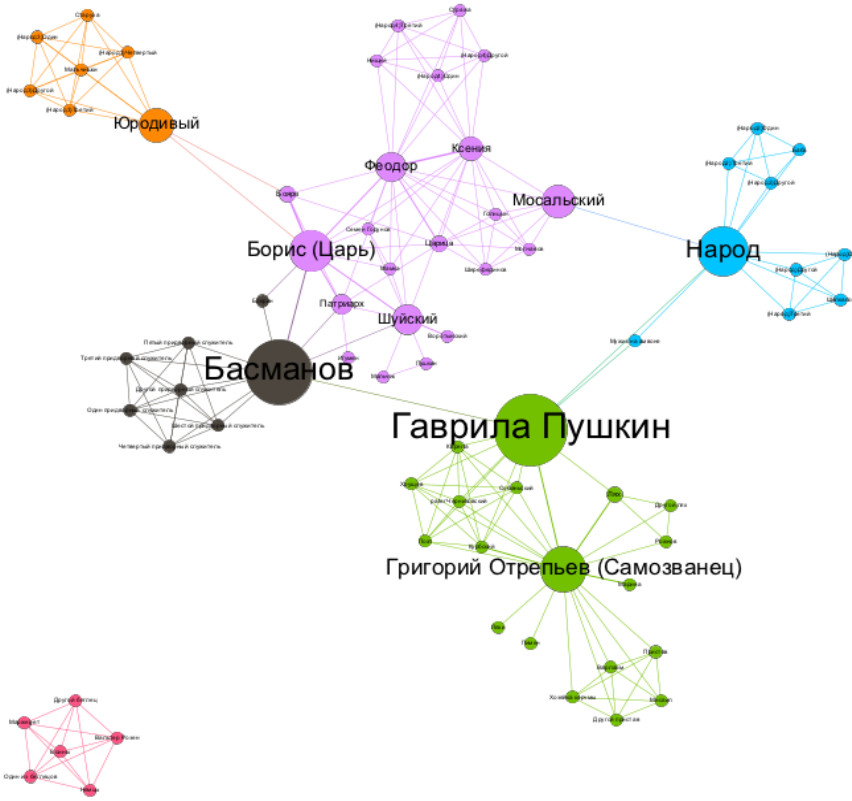


Figure 14. Boris Godunov network, nodes proportional to betweenness centrality

As one may notice, the most central character now is Gavrila Pushkin—clearly not one of the main heroes of the play. However, his high betweenness centrality is fairly obvious to those familiar with the plot. Gavrila Pushkin (one of the two Pushkin characters in the play) acts as a messenger ad mediator: he is being sent from Poland to Moscow to convey the False Dimitry's terms to Boris, and then he embarks on a mission to convince military chief Basmanov change sides—which eventually helps False Dimitry win the throne. After fulfilling this task, Gavrila Puskin as a follower of Dimitry, announces the decrees of the new tsar to the People ('Народ'), thus becoming the character that connects all cluster in the play.

This leads us to think that network metrics can sometimes reflect specific *functions* of characters in plays. And the function of Gavrila Pushkin may not be an *accidental* one, but rather deliberate: the idea that Pushkin's noble ancestors were actively *involved* in the Russian history, and especially the history of the Time of Troubles (Смутное время) can be traced throughout Pushkin's lyrics—see, for example, his famous poem 'My Pedigree' ("Мы к оной руку приложили"). All in all, such findings foreshadow new opportunities to network-oriented research of character systems in fiction.

Conclusions

In this paper we presented an open research-oriented corpus of Russian drama suitable for large-scale literary studies. Although the corpus is at the early stage of development, it can already serve as the basis for diverse research on structure and structural evolution of Russian drama, as we strove to highlight in our three cases studies above.

Later on, we hope to add more layers of annotation (e.g. named entities, classes of stage directions) and metadata (genres of the plays, social statuses of the characters etc.). This will open up new. The availability of compatible non-Russian corpora with similar markup obviously calls for cross-cultural research. RusDraCor is released under a free license, se we welcome derivation and enrichment efforts from third parties.

Acknowledgements

The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2017–2018 (grant № 17-05-0054) and by the Russian Academic Excellence Project “5-100”.

References

1. *Бертольт Брехт.* Теория эпического театра // Бертольт Брехт. Театр. Пьесы. Статьи. Высказывания. В пяти томах. Т. 5/2 М., Искусство, 1965
2. *Лотман Ю. М.* Структура художественного текста // Лотман Ю. М. Об искусстве. — СПб.: «Искусство — СПБ», 1998. — С. 14–285.
3. *Сапогов В. А.* Некоторые характеристики драматургического построения комедии А. Н. Островского «Лес» // А. Н. Островский и русская литература. Кострома, 1974
4. *Ярхо Б. И.* Распределение речи в пятиактной трагедии: (К вопросу о классицизме и романтизме): Подгот. текста, публ. и примеч. М. В. Акимовой; Предисл. М. И. Шапира // Philologica, 1997, т. 4, № 8/10, 201–284.
5. *Agarwal A., Kotalwar A., Rambow O.* (2013). Automatic Extraction of Social Networks from Literary Text: A Case Study on Alice in Wonderland, In Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013), Nagoya, Japan.
6. *Agarwal A., Corvalan A., Jensen J., Rambow O.* (2012). Social network analysis of Alice in Wonderland. In Proceedings of the NAACL HLT 2012 Workshop on Computational Linguistics for Literature, 88–96, Montreal, Canada.
7. *Alberich, R., Miro-Julia, J., Rossello, F.* (2002). Marvel universe looks almost like a real social network. Available at: https://arxiv.org/abs/cond-mat/0202174 (accessed December 29, 2017)
8. *Ardanuy M., Sporleder C.* (2014). Structure-based clustering of novels. In Proceedings of the EACL Workshop on Computational Linguistics for Literature, 31–39.
9. *Bastian M., Heymann S., Jacomy M.* (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

10. *Blondel V. D., Guillaume J., Lambiotte R., Lefebvre E.* (2008). Fast unfolding of communities in large networks, In Journal of Statistical Mechanics: Theory and Experiment, 10, 1000.
11. *Burrows, J. F.* (2007). All the way through: testing for authorship in different frequency strata. Literary and Linguistic Computing, 22(1): 27–48.
12. *Burrows, J. F.* (1987) Computation into Criticism: A Study of Jane Austen's Novels. Oxford. Clarendod Press.
13. *Bodrova A., Bocharov V.* (2014). Relationship Extraction from Literary Fiction. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2014". Available at: http://www.dialog-21.ru/digests/dialog2014/materials/pdf/BodrovaAABocharovVV.pdf (accessed December 29, 2017)
14. *Craig, H. and Kinney, A. F.* (2009). Shakespeare, Computers, and the Mystery of Authorship. Cambridge: Cambridge University Press.
15. *Dessen A.* (2011). Stage Directions and the Theater Historian. In The Oxford Handbook of Early Modern Theatre. : Oxford University Press.
16. *Dryden, John* (1668). Jack Lynch (Ed.), ed. An Essay of Dramatick Poesie. Available at: http://andromeda.rutgers.edu/~jlynch/Texts/drampoet.html
17. *Eder, M., Rybicki, J. and Kestemont, M.* (2016). Stylometry with R: a package for computational text analysis. R Journal, 8(1): 107–121, url: https://journal.r-project.org/archive/2016/RJ-2016-007/index.html
18. *Elson, D. K., Dames, N. and McKeown, K.* (2010). Extracting Social Networks from Literary Fiction, In Proceedings of ACL 2010, Uppsala, Sweden.
19. *Fischer F., Göbel M., Kampkaspar D., Kittel C., Trilcke P.* (2017). Network Dynamics, Plot Analysis: Approaching the Progressive Structuration of Literary Texts, In Digital Humanities 2017 Book of Abstracts. Montréal: McGill University
20. *Fischer, F.; Göbel, M.; Kampkaspar, D.; Trilcke, P.:* (2016a)Theatre Plays as 'Small Worlds'? Network Data on the History and Typology of German Drama, 1730–1930. DH2016, Kraków. URL: http://dh2016.adho.org/abstracts/360.
21. *Fischer, F., Vogel, A., Göbel, M., Kampkaspar, D., Trilcke, P.* (2016b) Distant-Reading-Showcase: 200 Jahre deutscher Dramengeschichte auf einen Blick. In: Digital Humanities im deutschsprachigen Raum (DHd) 2016 Konferenzabstracts
22. *Gleiser. P. M.* (2007). How to become a superhero. In Journal of Statistical Mechanics: Theory and Experiment, 9
23. *Glorieux, F.* (2016) Dramagraphie 0.2. Online source, April 4th, 2016. URL: http://resultats.hypotheses.org/749.
24. *Grayson S., Wade K., Meaney G., Greene D.* (2016) The Sense and Sensibility of Different Sliding Windows in Constructing Co-occurrence Networks from Literature. In: Bozic B., Mendel-Gleason G., Debruyne C., O'Sullivan D. (eds) Computational History and Data-Driven Humanities. CHDDH 2016. IFIP Advances in Information and Communication Technology, vol 482. Springer, Cham. DOI: https://doi.org/10.1007/978-3-319-46224-0_7
25. *Lee J., Wong T.* (2016). Conversational Network in the Chinese Buddhist Canon. In Open Linguistics 2016, 2, 427–436. DOI 10.1515/opli-2016–0022

26. *Lee J., Yeung C. Y.* (2012). Extracting Networks of People and Places from Literary Texts. In Proceedings of 26th Pacific Asia Conference on Language, Information, and Computation (PACLIC). 209–218
27. *Moretti F.* (2013), Distant Reading. Verso, London
28. *Moretti F.* (2011). Network Theory, Plot Analysis. In Stanford Literary Lab Pamphlets, Stanford, CA. Available at: https://litlab.stanford.edu/LiteraryLabPamphlet2.pdf (accessed December 29, 2017)
29. *Moretti F.* (2009). Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850). In: Critical Inquiry, Vol. 36, No. 1 (Autumn 2009), pp. 134–158
30. *Mueller M.* (2014). Shakespeare His Contemporaries: Collaborative Curation and Exploration of Early Modern Drama in a Digital Environment. Digital Humanities Quarterly. 8.3.
31. *Schofield A., Mehr L.* (2016). Gender-Distinguishing Features in Film Dialogue. In Proceedings of NAACL Workshop on Computational Linguistics for Literature 2016
32. *Schweizer T., Schnegg M.* (1998). The social structure of Simple Stories: network analysis [Die soziale Struktur der. „Simple Storys": Eine Netzwerkanalyse]. Available at: https://www.ethnologie.uni-hamburg.de/pdfs-de/michael-schnegg/simple-stories-publikation-michael-schnegg.pdf (accessed December 29, 2017)
33. *Schöch, C.; Henny, U.; Calvo Tello, J.*(2017) cligs/textbox: Spring is coming release. Data set, Zenodo, March 10th, 2017. URL: http://doi.org/10.5281/zenodo.376666.
34. *Skorinkin D.* (2017) Extracting Character Networks to Explore Literary Plot Dynamics, in: Computational Linguistics and Intellectual Technologies: papers from the Annual conference 'Dialogue '(Moscow, may 31—june 3 2017 r.). Issue. 16 (23): ed.: V. Selegey. V. 1. M.: RSUH, 2017. pp. 257–270.
35. *Stiller J., Hudson M.* (2005). Weak Links and Scene Cliques Within the Small World of Shakespeare. In Journal of Cultural and Evolutionary Psychology 3, no. 1
36. *Stiller J., Nettle D., Dunbar R.* (2003). The Small World of Shakespeare's Plays. In Human Nature, 14(4), 397–408
37. *Trilcke P., Fischer F., Göbel M., Kampkaspar D.* (2015a), Comedy vs. Tragedy: Network Values by Genre. Network Analysis of Dramatic Texts. Available at: https://dlina.github.io/Network-Values-by-Genre/ (accessed December 29, 2017)
38. *Trilcke P., Fischer F., Kampkaspar D.* (2015b). Digitale Netzwerkanalyse dramatischer Texte, In DHd2015. Von Daten zu Erkenntnissen 23. bis 27. Graz. Book of Abstracts. Austrian Centre for Digital Humanities.
39. *Trilcke P., Fischer F., Göbel M., Kampkaspar D.* (2016). Theatre Plays as 'Small Worlds'? Network Data on the History and Typology of German Drama, 1730–1930. In Digital Humanities 2016: Conference Abstracts. Jagiellonian University, Pedagogical University, Kraków, 385–387.
40. *Weitin T.* (2017), Scalable Reading. Zeitschrift für Literaturwissenschaft und Linguistik, Volume 47, Issue 1, pp 1–6
41. *Xanthos, A. et al.* (20016) Visualising the Dynamics of Character Networks. Proceedings, DH2016. Jagiellonian University & Pedagogical University, Kraków, 417–419.

АНАЛИЗ РЕЧЕВЫХ СБОЕВ В ДИСКУРСЕ РУССКОЯЗЫЧНЫХ ДЕТЕЙ 10–12 ЛЕТ[1]

Слабодкина Т. А. (slabodkina.t@gmail.com)
МГУ имени М. В. Ломоносова, Москва, Россия

Федорова О. В. (olga.fedorova@msu.ru)
МГУ имени М. В. Ломоносова,
Институт языкознания РАН, РАНХиГС, Москва, Россия

Данная работа продолжает уже ставшую традиционной для конференций «Диалог» проблематику исследования речевых сбоев (см., в частности, работы Подлесская, Комарова 2010; Лауринавичюте, Федорова 2010; Подлесская 2013; Богданова-Бегларян 2013; Подлесская 2014; Потанина и др. 2016). В настоящей статье этот вопрос будет рассмотрен при сравнении языкового поведения русскоязычных детей 10–12 лет (**раздел 1**) со взрослыми носителями языка на материале корпуса тангramм (**раздел 2**). В **разделе 3** будет приведена классификация речевых сбоев, в **разделе 4** приведены результаты исследования. Наконец, **раздел 5** будет посвящен обсуждению результатов и перспективам дальнейшей работы. Мы покажем, что дискурсивное поведение ребенка 10–12 лет с точки зрения речевых сбоев отличается от аналогичного поведения взрослых носителей, что подтверждает нашу гипотезу о позднем дискурсивном развитии ребенка.

Ключевые слова: диалог, онтогенез, тангramмы, порождение речи, речевые сбои

[1] Работа выполнена при финансовой поддержке РФФИ, грант № 17-06-00264 «Взаимосвязь между формированием дискурсивных и когнитивных навыков русскоязычных детей 11–12 лет».

SPEECH DISFLUENCIES ANALYSIS IN THE DISCOURSE OF 10–12 YEARS OLD NATIVE RUSSIAN SPEAKING CHILDREN

Slabodkina T. A. (slabodkina.t@gmail.com)
Lomonosov Moscow State University, Moscow, Russia

Fedorova O. V. (olga.fedorova@msu.ru)
Lomonosov Moscow State University, Institute of Linguistics RAS, RANEPA, Moscow, Russia

The paper reviews the problem of speech disfluency which over the years has become traditional for the "Dialogue" conference (see Podlesskaya, Komarova 2010; Laurinavichyute, Fedorova 2010; Fedorova 2010; Podlesskaya 2013; Bogdanova-Beglarian 2013; Podlesskaya 2014; Potanina et al. 2016).

In this paper, we compared speech disfluencies in two corpora of dialogues between children of 10–12 years old (**section 1**) and adults (**section 2**). Both corpora were collected using the referential communication task "Tangrams" (to perform the task, participants had to agree on the nomination of some abstract figures).

In the **third section** of the text, the authors provide the classifications of speech disfluencies present in the dialogues with examples. The results of the comparison and the methods of analysis are given in the **fourth paragraph**. Finally, the **last section** contains the discussion of the results and perspectives of the further work. The paper shows that speech of children of the given age group differs from adults' speech in terms of disfluencies at the discourse level.

Key words: dialogue, ontogeny, tangrams, speech production, speech disfluency

1. Языковое развитие русскоязычных детей 10–12 лет

Лингвистика детской речи занимается по большей части исследованием речи ребенка в возрасте от одного года до пяти лет, так как именно в этот период происходит становление базовых языковых функций. Отечественная онтолингвистика (термин С. Н. Цейтлин, см., напр., [Цейтлин 2000]) часто идет дальше и рассматривает языковое развитие ребенка вплоть до 8–9 лет. Однако, как западная, так и отечественная наука пока уделяют мало внимания исследованию речи детей 11–12 лет (в качестве редкого исключения можно отметить учебник [Hoff 2005]). Между тем этот возраст — вступление в пору отрочества — представляет собой важный языковой феномен, особенно в том, что касается формирования дискурсивных и коммуникативных навыков устной речи (о формировании дискурсивных и коммуникативных навыков детей до пяти

лет, см., в частности, известную работу [Pan, Snow 1999]). В целом, устная речь является первичной и наиболее важной формой общения между людьми, однако в настоящее время ее исследованию уделяется еще недостаточно внимания, как в области «взрослой» лингвистики, так и в области усвоения языка.

Говоря о формировании дискурсивных навыков русскоязычного ребенка, в первую очередь стоит упомянуть работы К. Ф. Седова, который проводил много разнообразных исследований речи школьников всех возрастов [Седов 2004]. В частности, К. Ф. Седов оспаривает точку зрения о раннем референциальном развитии ребенка к 4 годам [Matthews et al. 2009] и утверждает, что к шести–семи годам «человек становится обладателем языкового механизма, своего рода персонального компьютера, который открывает перед ним новые когнитивно-коммуникативные возможности. Однако получив в свое распоряжение языковой механизм, ребенок еще не имеет навыков использования его в речевой деятельности, у него как бы еще нет компьютерного программного обеспечения. Использовать язык ребенок учится в повседневном общении, в каждодневной речевой практике, которая состоит из порождения и смыслового восприятия многочисленных речевых произведений» [Седов 2004: 20–21]. Седов пишет, что становление продолжается и в школьном возрасте, когда ребенок постепенно повышает уровень своей дискурсивной компетенции, а полное овладение происходит только в старшем подростковом возрасте (к 15–16 годам). Так Седов формулирует гипотезу о позднем референциальном развитии. В данной работе мы выдвигаем более широкую гипотезу — ***о позднем дискурсивном развитии ребенка***. Мы предполагаем, таким образом, что дискурсивное поведение ребенка 10–12 лет как в целом, так и с точки зрения речевых сбоев, будет отличаться от аналогичного поведения взрослых.

2. Метод референциальной коммуникации «Танграммы»

Настоящее исследование было проведено на материале диалогического корпуса описаний танграмм — фигурок из китайской игры-головоломки (см. рис. 1), которые не имеют устоявшихся вербальных дескрипций. Насколько нам известно, в научных целях танграммы были впервые использованы в исследовании, описанном в работе Clark, Wilkes-Gibbs 1986[2]. С тех пор было проведено множество разнообразных исследований [см. обзор в Федорова 2014], в том числе и несколько работ на русском материале [Зинова и др. 2011; Федорова и др. 2013; Федорова 2017].

[2] Как пишет сам Кларк, идею использовать китайские танграммы ему подсказала А. В. Беляева во время его визита в Россию. Однако Кларк был не первым — согласно В. Лефельту, впервые танграммы были использованы в научных целях Э. Эспером еще в 1933 г. [Levelt 2013].

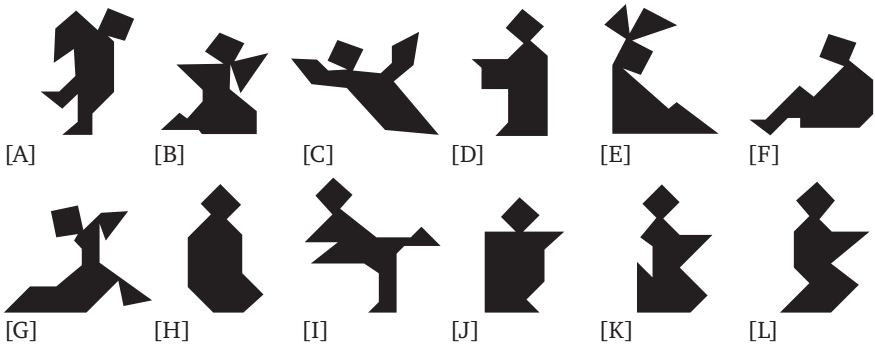


Рис. 1. Стимульный материал исследования из пионерской работы Clark, Wilkes-Gibbs 1986 (в ходе эксперимента на карточках не было буквенных обозначений)

Основная идея метода ***референциальной коммуникации*** (Referential communication task)[3], введенный в практику в 70-ые годы XX века социальным психологом Р. Краусcом [Krauss, Weinheimer 1966], состоит в том, что один из собеседников, Инструктор (Director) видит или знает нечто, что он должен вербально передать второму собеседнику, Раскладчику (Matcher), который этого не видит или не знает. Одну из разновидностей этого метода — методику «беспорядок» — и усовершенствовал Г. Кларк. Усовершенствованная им методика заключается в следующем. Участники эксперимента сидят за столами друг напротив друга, между ними находится перегородка, благодаря которой они не могут видеть карточки, которые располагаются на столе их собеседника. Перед каждым испытуемым лежит одинаковый набор из 12 карточек с черной картинкой-силуэтом (танграммой), при этом у одного из них картинки разложены в заранее определенном порядке, а у второго карточки с теми же картинками перемешаны. Первый участник — Инструктор — должен объяснить правильное расположение танграмм второму участнику — Раскладчику, по очереди описывая каждую карточку.

В нашем исследовании с русскоязычными испытуемыми, у каждой пары испытуемых было четыре попытки, в ходе выполнения заданий они не менялись ролями. В исследовании приняли участие 36 пар студентов МГУ имени М. В. Ломоносова. Собранный корпус состоит из 63 тысяч слов и 8500 реплик, общая длительность около 10 часов; для более подробного изучения и проведения настоящего сравнения с данными «детского» исследования из собранного корпуса случайным образом были выбраны 16 записей. Кроме того, было проведено исследование с 18 парами детей возраста 10–12 лет[4], из которых для нашего исследования по техническим причинам были отобраны 16 за-

[3] Мы будем использовать дословный перевод английского термина «referential communication», другой принятый перевод — «референтное общение» [Самойленко 2010].

[4] Авторы выражают благодарность учителю русского языка и литературы лицея № 1564 Н. Н. Ципенко за неоценимую помощь в организации эксперимента с детьми.

писей. В итоге анализируемый корпус, включающий взрослые и детские записи, состоит из 32 диалогов общей длительностью 6 часов 37 минут (каждый длительностью от 4 мин. 32 с. до 25 мин.).

3. Классификация речевых сбоев

Как известно, речевые сбои могут иметь разную природу, в том числе они маркируют и затруднения в речепорождении. Мы предположили, что при сравнении диалогической речи детей и взрослых материал первых будет иметь значимо больше речевых сбоев, свидетельствующих о большем количестве затруднений. Все записанные диалоги были расшифрованы и размечены с целью дальнейшего анализа речевых сбоев. В ходе аннотирования мы опирались на правила, разработанные при анализе спонтанной устной речи в [Кибрик, Подлесская 2009], [Подлесская 2014] и [Потанина и др. 2016]. Среди явлений, которые могут быть названы речевым сбоем или свидетельствовать о затруднении, мы выделяем следующие:

1. **Паузы.** С одной стороны, паузы обозначают места хезитации и временных затруднений при формулировании высказывания; с другой стороны, они являются неотъемлемой частью любого устного дискурса и помогают не только говорящему, предоставляя дополнительное время для обдумывания дальнейшего поведения, но и слушающему, организуя речевые отрезки устной речи. В нашем корпусе мы отмечали два основных вида пауз — абсолютные и заполненные («экания» и «мекания»), а также удлинение гласных внутри слов при хезитации. Абсолютные паузы отмечались точками, а заполненные имели буквенное обозначение в зависимости от характера звука (*э, а* или *м*). Все паузы также были измерены и поделены на группы по длине (в нашей разметке графически различаются количеством точек).

(1) *Р: Последняя картинка ···(0.6) это···(0.7) человек, изображённый вполовину ····(1.2) ээ(0.3) выставил руки вперёд и вверх*

2. **Обрыв слова.** Обрывы слова часто являются сигналом речевого затруднения и сочетается с другими видами сбоев, в нашем корпусе обозначается знаком «=»:

(2) *Я: ····(1) э но ноги у него прям яв=|| как|| как настоящие да?*

3. **Самоисправления.** Пример 3 демонстрирует образец самоисправления (обозначается «||»):

(3) *К: [да и у нее одна] как бы она ···(0.6) не-ет это не то {ЦОКАНЬЕ} вот и у нее одна рука ····(1.2) треугольная {СМЕХ} она отс=|| присоединена только к шее*

В этом примере фрагмент *отс=* как бы «зачеркивается» говорящим (что совпадает с обрывом слова и, если его убрать, получается полноценное предложение *она присоединена к шее*). В данном случае «устраняется» лишь фрагмент

реплики, структура единицы принципиально не меняется (такие случаи мы обозначаем знаком «||», к ним же мы относили и повторы).

В примере (4), напротив, говорящий отказывается от первоначального плана построения и отбраковывает запланированное развитие реплики полностью:

(4) *Я: ага ага он== ···(0.6) да вроде поняла ···(0.5) давай дальше*

В примере (5) говорящий как бы возвращается к завершенной единице *на ногах лежит* и исправляет, редактирует ее: *ну не лежит а сидит на них.*

(5) *К: под головой как бы еще один ромбик · · · ·(4.5) а потом внизу · · · ·(1.8) ммм(0.6) ноги · · · ·(1.0) ну не совсем на коленках стоит он · · (0.4) на ногах лежит · · · ·(2.8) ну не лежит а сидит на них*

4. **Лексические маркеры речевых сбоев, или «маркеры эмоциональной реакции на речевую проблему»** [Потанина и др. 2016]. В приводимом ниже примере междометие *ой* эксплицитно свидетельствует о том, что говорящий признает ошибку или затруднение:

(6) *К: Первая картинка в первом ряду ···(0.5) в верхнем этоаа ····(1.3) человек который повернул голову на триста шестьдесят градусов ой назад {ОТКАШЛИВАНИЕ}*

5. **Маркеры препаративной подстановки**. Маркерами препаративной подстановки называются такие слова, которые «подставляются» говорящим, пока происходит подбор слова, повторяющие его грамматические характеристики и интонационный контур, иногда также реализуются с нисходящим акцентом или без акцента.

(7) *И: на следующей картинке человек ээ(0.4) ···(0.8) ээ(0.2) п=|| \<u>такой</u> \ присевший ··(0.3) ээ(0.2) коленки у него согнуты, и-и руки смотрят вперёд.*

В корпусе также были размечены невербальные явления речи, такие как смех, кашель, свист, вздох и др.

4. Результаты исследования

4.1. Сравнение темпа речи

В нашей работе под темпом речи мы будем понимать количество слов в минуту (полноценных и отбракованных в последствии говорящим). Мы подсчитали количество слов в минуту для каждого диалога (см. **таблицу 1**) и сравнили «детскую» и «взрослую» выборки статистическим методом с помощью теста Манна-Уитни, подходящим для обработки независимых выборок небольшого объема, в программе STATISTICA. В результате мы получили уровень значимости <0,05, что свидетельствует о наличии достоверной разницы

по рассматриваемому параметру. Темп речи взрослых испытуемых оказался значимо выше, чем темп речи детей. Этот вывод важен как сам по себе, так и для дальнейшего подсчета количества речевых сбоев, которые мы должны рассматривать не на единицу времени, а на 100 слов.

Таблица 1. Сравнение темпа речи взрослых испытуемых и детей

диалог №	Количество слов в минуту	
	Взрослые	Дети 10–12 лет
1	183,8	154,9
2	132,6	71,9
3	101,0	60,7
4	123,3	104,9
5	192,6	71,0
6	148,3	62,4
7	139,3	73,9
8	146,8	90,1
9	155,4	134,7
10	117,6	64,7
11	232,9	178,3
12	110,6	118,8
13	130,3	106,1
14	109,2	109,7
15	138,0	74,3
16	72,9	74,0

4.2. Сравнение количества речевых сбоев

На следующем этапе мы определили количество речевых сбоев на каждые 100 слов в диалогическом общении отдельно для взрослых и для детей, см. **таблицу 2**. Сравним полученные данные с данными из работы [Потанина и др. 2016], согласно которой количество речевых сбоев в монологической речи взрослых русских носителей колеблется в пределах от 0,77 до 8,58 на 100 слов (среднее 4,57), что в целом несколько выше средних значений, характерных для спонтанной речи (что для английских устных пересказов составляет 1,9–3,7 на 100 слов [Fraundorf, Watson 2008], для японских монологов — 1,2 на 100 слов [Maruyama, Sano 2006], для спонтанных диалогов в иорданском варианте арабского языка — 1,6 на 100 слов [Al-Harahsheh 2015]. В наших данных количество речевых сбоев оказалось на порядок больше (среднее для взрослых 22,28 на 100 слов, для детей — 33,41 на 100 слов), так как кроме лексико-грамматических маркеров речевых сбоев в рассмотрение были включены также все паузы хезитации и фонологически не мотивированные удлинения звуков; кроме того, известно, что диалогическая речь содержит больше речевых сбоев, чем монологическая.

Таблица 2. Количество сбоев на каждые 100 слов во «взрослом» и «детском» корпусе (для удобства чтения числа округлены до десятых)

Взрослые					Дети 10–12 лет				
Количество слов	Общее количество сбоев	Количество пауз на 100 слов	Количество самоисправлений на 100 слов	Общее количество сбоев на 100 слов	Количество слов	Общее количество сбоев	Количество пауз на 100 слов	Количество самоисправлений на 100 слов	Общее количество сбоев на 100 слов
2070	296	11,7	2,6	14,3	1894	385	16,3	4,1	20,3
1505	328	20,1	1,7	21,8	1798	630	32,2	2,8	35,0
1404	472	29,6	4,1	33,6	1113	529	44,9	2,6	47,5
916	292	27,2	4,7	31,9	1468	599	35,4	5,4	40,8
3532	706	14,6	5,4	20,0	1315	490	31,0	6,2	37,3
1194	211	14,3	3,4	17,7	705	197	25,1	2,8	27,9
850	176	18,4	2,4	20,7	999	450	40,3	4,7	45,0
1094	204	13,1	5,6	18,6	1636	451	24,2	3,4	27,6
1125	291	18,3	7,6	25,9	2489	528	14,8	6,4	21,2
944	226	20,3	3,6	23,9	1102	544	42,5	6,9	49,4
1938	179	5,3	4,0	9,2	4133	748	11,0	7,1	18,1
667	156	19,9	3,4	23,4	1251	337	17,4	9,5	26,9
563	161	24,5	4,1	28,6	1506	443	25,0	4,4	29,4
827	259	28,3	3,0	31,3	556	117	18,0	3,1	21,0
1521	203	10,5	2,8	13,3	1226	660	46,7	7,2	53,8
894	298	28,3	5,0	33,3	903	373	16,3	3,0	41,3

Посмотрим теперь, отличается ли количество речевых сбоев на каждые 100 слов в диалогическом общении взрослых и детей. Статистический анализ (в этом случае для сравнения данных мы также применили критерий Манна-Уитни) демонстрирует значимое различие между двумя независимыми группами ($p < 0,05$), что подтверждает нашу гипотезу о большем количестве маркеров затруднений в речи детей 10–12 лет.

5. Обсуждение результатов и перспективы дальнейшей работы

В настоящее время когнитивные и языковые навыки 11–12-летних детей исследованы недостаточно полно. Данный возраст, таким образом, попадает в некоторую «возрастную яму» между, с одной стороны, относительно хорошо изученными когнитивными и языковыми навыками детей дошкольного и младшего школьного возраста и когнитивными и дискурсивными навыками взрослых людей, с другой стороны. Данная работа в некоторой степени восполняет этот пробел и демонстрирует значимые отличия речи детей данной

возрастной группы от речи взрослых. Более конкретно, темп речи детей в описанных диалогах оказывается значимо ниже, а количество речевых сбоев значимо больше, чем в аналогичных диалогах между взрослыми. Существуют ли другие особенности и какие именно — вопрос для дальнейших исследований в этой области.

В дальнейшей работе с собранным корпусом танграмм возможно более подробное изучение и сравнение отдельных видов речевых сбоев взрослых и детей и анализ их природы. В частности, в работе [Fraundorf, Watson 2013] авторы обнаружили, что трем разным типам речевых сбоев — незаполненным паузам, заполненным паузам и повторам слова — соответствуют разные типы трудностей при порождении речи.

Также интерес представляют специфические для диалога сбои — например, связанные с наложением реплик говорящих. Кроме того, отдельного внимания заслуживает вопрос, как реагируют взрослые и дети на перебивание — продолжают ли они свою реплику или, наоборот, обрывают. Воспринимается ли перебивание как оптимизация усилий и экономия времени или как помеха и отвлекающий фактор? В целом, в дальнейшей работе предстоит выяснить, какие правила построения диалога у ребенка уже сформировались к 10–12 годам, а какие еще находятся в процессе усвоения.

Литература

1. *Al-Harahsheh A. M. A.* (2015), A Conversation Analysis of self-initiated repair structures in Jordanian Spoken Arabic, Discourse Studies, Vol. 17 (4), pp. 397–414.
2. *Bogdanova-Beglarian N. V.* (2013), Those who seek, will they find? (search function of verbal hesitations in Russian spontaneous speech) [Kto ishchet — vsegda li najd'ot? (o poiskovoj funkcii verbal'nykh khezita-tivov russkoj spontannoj rechi)]. Computational Linguistics and Intellectual Technologies: papers from the Annual International Conference "Dialogue". V. 12 (19). Moscow: RSUH, pp. 125–136.
3. *Clark H. H., Wilkes-Gibbs D.* (1986), Referring as a collaborative process, Cognition, Vol. 22(1), pp. 1–39.
4. *Fedorova O. V., Delikishkina E. A., Slabodkina T. A., Tsipenko A. A.* (2013), Dialogue modelling in psycholinguistics: literal and analogical perspectives as bases for adults' and children's references [Modelirovanie dialoga v psikholingvistike vzroslye i detskie strategii opisaniya obektov deystvitelnosti], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2013" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2013"], RGGU, Moscow, pp. 230–238.
5. *Fedorova O. V.* (2014), Experimental discourse analysis [Eksperimentalnyy analiz diskursa]. Yazyki Slavyanskoy Kultury, Moscow.
6. *Fedorova O. V.* (2017) The Contribution of the Kinetic Component to the Multimodal Communication, or Tangram Description Strategies Revisited. [Vklad kineticheskoj sostavlyayushhej v mul'timodal'nuyu kommunikatsiyu, ili Eshhe raz

o strategiyakh opisaniya tangramm]. Computational linguistics and intellectual technologies: proceedings of the annual international conference "Dialogue". V. 16, pp. 118–133.
7. *Fraundorf S. H., Watson D. G.* (2008), Dimensions of variation in disfluency production in discourse, in J. Ginzburg, P. Healey, Y. Sato (Eds.), Proceedings of LONDIAL 2008, the 12th Workshop on the Semantics and Pragmatics of Dialogue, London: King's College London, pp. 131–138.
8. *Fraundorf S. H., Watson D. G.* (2013), Alice's adventures in um-derland: Psycholinguistic dimensions of variation in disfluency production, Language, Cognition and Neuroscience, Vol. 29, pp. 1083–1096.
9. *Hoff E.* (2005), Language Development. Belmont, CA, Wadsworth/Thomson Learning.
10. *Kibrik A. A., Podlesskaya V. I. [Eds.]* (2009), Rasskazy o snovidenijax: korpusnoe issledovanie usntogo russkogo diskursa [Night Dream Stories: A corpus study of spoken Russian discourse]. Moskva: Jazyki Slavjanskix Kul'tur.
11. *Krauss R. M., Weinheimer S.* (1966), Concurrent feedback, confirmation, and the encoding of referents in verbal communication, Journal of Personality and Social Psychology, Vol. 4(3), pp. 343–346.
12. *Laurinavichyute A. K., Fedorova O. V.* (2010), Effects of hesitation in speech on syntactic structure in comprehension: evidence from russian speakers [Vliyanie pauzy khezitatsii na ponimanie sintaksicheskoj struktury predlozheniya nositelyami russkogo yazyka]. Computational linguistics and intellectual technologies: papers from the annual international conference "Dialogue". V. 9 (16). Moscow: RSUH, pp. 279–283.
13. *Levelt W. J. M.* (2013), A History of Psycholinguistics: The pre–Chomskyan era, Oxford University Press, Oxford.
14. *Maruyama T., Sano Sh.* (2006), Classification and Annotation of Self-Repairs in Japanese Spontaneous Monologues, LPSS — Linguistic Patterns in Spontaneous Speech, Taipei, pp. 283–298.
15. *Matthews D. E., Lieven E. V. M., Tomasello M.* (2009), The Development of Reference from Two to Four Years, Proceedings of the "Production of Referring Expressions 2009: Bridging the gap between computational and empirical approaches to reference" conference, Amsterdam.
16. *Pan B., Snow C.* (1999), The development of conversational and discourse skills, M. D. Barrett (ed.) The Development of Language, London: Psychology Press, pp. 229–250.
17. *Podlesskaya V. I., Komarova A. D.* (2010), Self-repairs in Japanese narrative discourse: a corpus-based case-study [Samoispravleniya govoryashhego v yaponskom ustnom narrative: analiz korpusnykh dannykh] Computational linguistics and intellectual technologies: papers from the annual international conference "Dialogue". V. 9 (16). Moscow: RSUH, pp. 382–388.
18. *Podlesskaya V. I.* (2013), Vague reference in Russian: evidence from spoken corpora [Nechetkaja nominacija v russkoj razgovornoj rechi: opyt korpusnogo issledovanija]. Computational Linguistics and Intellectual Technologies: papers from the Annual International Conference "Dialogue". V. 12 (19). RSUH, Moscow, pp. 561–573.

19. *Podlesskaya V. I.* (2014), They shot him dead, oh, no, they knifed him dead with a saber: self-repairs in oral stories [To est', ne ubili, a zarezali sablej: samoispravlenija govorjashhego v ustnyh rasskazah]. Computational Linguistics and Intellectual Technologies: papers from the Annual International Conference "Dialogue". V. 13 (20). Moscow: RSUH, 2014, pp. 526–540.
20. *Potanina Y. D., Podlesskaya V. I., Fedorova O. V.* (2016), Verbal Working Memory and Speech Production Difficulties: Data from Russian Multimodal Corpus [Verbal'naya rabochaya pamyat' i leksiko-grammaticheskie signaly rechevykh zatrudnenij: dannye russkogo mul'timodal'nogo korpusa] Computational linguistics and intellectual technologies: proceedings of the annual international conference "Dialogue". V. 15 (22), Moscow: RSUH, p.566–576.
21. *Samoilenko E. S.* (2010), Problems of comparison in psychological research [Problemy sravneniya v psikhologicheskom issledovanii]. Publishing House of the IP, RAS, Moscow.
22. *Sedov K. F.* (2004), Discourse and personality [Diskurs i lichnost'], Moscow, Labirint.
23. *Tsejtlin S. N.* (2000), Language and child. Liguistics of children's speech. [Yazyk i rebenok. Lingvistika detskoj rechi]. Valdos, Moscow.
24. *Zinova Ju. A., Dragoy O. V., Fedorova O. V.* (2011), Experimental study of verbal interaction: language pathology data [Eksperimentalnoe issledovanie rechevogo vzaimodeystviya: dannye yazykovoy patologii], Vestnik MGU, Ser. 9. Philology, 4, pp. 167–175.

GENDER, DECLENSION AND STEM-FINAL CONSONANTS: AN EXPERIMENTAL STUDY OF GENDER AGREEMENT IN RUSSIAN

Slioussar N. A. (slioussar@gmail.com)
Higher School of Economics, Moscow, and Saint-Petersburg State University, Saint-Petersburg, Russia

Every adult native speaker of Russian knows that *kon'* is masculine and *lan'* is feminine, although 3rd declension nouns present some difficulties in the first and second language acquisition. However, will the fact that these nouns are less frequent than masculine nouns ending in a consonant or feminine nouns ending in *-a/ja* play a role for online subject-predicate agreement processing? Or will subject-predicate agreement processing be more problematic with subjects of a certain gender? Finally, some final consonants are more characteristic for feminine gender, while the others for masculine gender. Are speakers sensitive to this? We present two experiments addressing these questions. We found that all three factors play a role, but for different tasks (online agreement processing or determining the gender of a novel word) and at different processing stages.

Keywords: grammatical gender, declension, experimental, Russian

РОД, СКЛОНЕНИЕ И КОНЕЧНЫЙ СОГЛАСНЫЙ ОСНОВЫ: ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ СОГЛАСОВАНИЯ ПО РОДУ В РУССКОМ ЯЗЫКЕ

Слюсарь Н. А. (slioussar@gmail.com)
НИУ ВШЭ, Москва, и СПбГУ, Санкт-Петербург, Россия

1. Introduction

The gender of many Russian nouns cannot be determined from their inflectional affixes. But nouns are not distributed evenly among genders and declensions. Firstly, masculine nouns are in general more frequent than feminine and neuter (48% vs. 35% and 17% forms in the grammatically disambiguated subcorpus of the Russian

National Corpus (RNC, http://www.ruscorpora.ru)). Secondly, a nominative singular form ending in *-a/ja* or with a zero inflection may be masculine or feminine, but the former is much more likely to be feminine, while the latter is much more likely to be masculine (more details about are given below). Thirdly, if the final consonant of a noun with a zero inflection is taken into account, we will see that, for example, most nouns ending in *-s'* are feminine, while most nouns ending in *-r'* are masculine.

In this paper, we present two experimental studies exploring whether native speakers of Russian are sensitive to the distributional properties outlined above. The first experiment is dedicated to the online processing of gender agreement. In the second pilot experiment, we study how participants determine the gender of real and nonce nouns.

Now let us discuss the relevant properties of nouns in more detail. Table 1 shows the distribution of nouns among genders and declensions in the grammatically disambiguated subcorpus of the RNC. Masculine nouns ending in *-a/ja* and feminine nouns with a zero inflection are termed non-prototypical due to their low relative frequency. These nouns are known to be problematic for the L1 and L2 acquisition (e.g. [Janssen, 2016]; [Rodina & Westergaard, 2012]; [Schwartz et al., 2015]; [Tseitlin 2000]). But [Rusakova 2013] who studied naturally occurring errors in spoken Russian found that adult native speakers do not make more gender agreement errors with such nouns[1]. Still, what does not show up in the number of errors may influence online sentence processing. Experiment 1 is dedicated to this question.

In this paper, we focus on consonant-final nouns. The distributional picture gets more complex if the nature of the final consonant is taken into account. The gender of the words ending in /ž/, /š/, /č'/, /š'č'/ can be determined orthographically, while words ending in other non-palatalized consonant can be only masculine. No such clues are available for the words ending in other palatalized consonants, but, as Table 2 shows, their distribution between the two genders varies greatly (to represent both the relative frequencies of different forms and the number of existing lemmas we relied not only on the RNC, but also on the *Grammatical Dictionary of the Russian Language* [GDRL, Zaliznjak, 1987]).

The majority of nouns ending in labials are feminine. Most nouns ending in *-r'* and *-l'* are masculine, many of them have agentive suffixes like *-tel'* or *-ar'*. Nevertheless, feminine nouns are not just singular cases in this group. Forms of nouns ending in other lingual consonants are either predominantly feminine or are evenly distributed between the two genders (but the number of feminine lemmas is still larger). The majority of nouns ending in *-t'*, which are especially numerous, have the suffix *-ost'*. In our pilot experiment 2, we explored whether adult native speakers are sensitive to these differences.

[1] Cases where gender variation is observed are not taken into account (a corpus-based study of such cases was conducted by [Savchuk 2011]).

Table 1. The distribution of nouns among genders and declensions in the grammatically disambiguated subcorpus of the RNC[2]

Declension and gender	Percentage of nouns in the RNC	Ending in Nom.Sg and prototypicality	Examples
1st decl. feminine	29% nouns	end in *-a/ja,* 'prototypical F'	*zhena* 'wife'
1st decl. masculine	1% nouns	end in *-a/ja,* 'non-prototypical M'	*djadja* 'uncle'
2nd decl. masculine	46% nouns	end in a consonant, 'prototypical M'	*syn* 'son', *gel'* 'gel'
2nd decl. neuter	18% nouns	end in *-o/e,* 'prototypical N'	*pole* 'field'
3rd decl. feminine	5% nouns	end in a consonant, 'non-prototypical F'	*mel'* 'shallow'
irregular and indeclinable	1% nouns		

Table 2. Nouns in the grammatically disambiguated subcorpus of the RNC and in the GDRL

	RNC (Nom.Sg forms)		GDRL (lemmas)	
Final consonant	M	F	M	F
/b'/	34 (24%)	110	1	11
/p'/	3 (2%)[3]	169	—	19
/v'/	13 (1%)	1,448	1	20
/f'/	0	2	—	2
/m'/	0	16	—	3
/d'/	748 (51%)	707	10	55
/t'/	713 (5%)	13,184	17	3414
/z'/	319 (49%)	327	7	34
/s'/	80 (14%)	491	5	57
/n'/	2,354 (45%)	2,842	126	112
/r'/	2,160 (76%)	677	177	34
/l'/	6,648 (71%)	2,653	1083	215

2. Previous experimental studies

Two groups of experimental studies are relevant for the present paper: analyzing gender agreement and nouns with more or less morphologically regular inflections.

[2] The counts are taken from [Slioussar and Samoilova 2015]. Substantivized adjectives were not taken into account.

[3] These three forms are *rup'* (a reduced form of the noun *rubl'* 'ruble').

There are relatively few experimental studies of gender agreement in Russian. In three of them ([Akhutina et al. 1999, 2001]; [Romanova & Gor 2017]) adjectives were presented before nouns audially or visually. In congruent conditions, adjectives agreed with the following nouns, in incongruent ones they did not, and some experiments also included a baseline condition where bare adjective stems without inflections or adverbs were presented. Several methods were employed, including lexical decision (answering whether the presented stimulus is a real word or a nonce word), grammaticality judgment (answering whether the presented fragment is grammatical) and cued-shadowing in which participants must repeat the second presented word (the target noun).

However, the question was always the same: would participants answer significantly faster and more accurately in congruent conditions compared to incongruent ones, and would there be any differences associated with the gender of the nouns? In experiments with a baseline condition, it was also possible to check whether the difference between congruent and incongruent conditions was primarily due to facilitation in the former, or to inhibition in the latter, or both effects were equally prominent. In brief, [Akhutina et al. 2001] observed significant facilitation and inhibition effects for feminine nouns, while for masculine nouns, only inhibition was significant, and for neuter ones, only facilitation was significant[4]. Results from other studies were similar.

The explanations offered in these studies go along the same lines. Masculine gender as the most frequent is assumed to be unmarked, or default, while neuter is considered the most marked. Thus, masculine is expected by default, and strengthening this expectation by a masculine adjective does not produce a big difference (hence no significant facilitation effects). Neuter is the least expected option, so priming a neuter noun with a neuter adjective has the largest effect compared to the baseline condition (hence facilitation effects for neuter nouns are larger than for feminine nouns). Inhibition effects are explained by rechecking, which is especially costly for masculine nouns presented after non-masculine adjectives.

None of these three studies looked at 3rd declension feminine nouns, while the experiments by [Taraban and Kempe 1999] specifically focused on them. Taraban and Kempe selected masculine and feminine nouns ending in a palatalized consonant (opaque condition) and in non-palatalized consonants or in *-a/ja*, which are unambiguously masculine or feminine (transparent condition). They examined the role of such transparency for subject–predicate agreement using word-by-word self-paced reading and forced choice tasks. Participants were asked to read sentence beginnings like (1a) or (2a) and then to select one of the two verb forms in the remaining fragment like (1b) or (2b). In some conditions, sentence fragments contained adjectives. Participants were adult native speakers and L2 learners. For native speakers, transparency and the presence of a gender-marked adjective did not play any role.

(1)	a.	*Daže*	*(obyčnaja)*	*muka/sol'*	*teper'…*
		even	ordinary$_{F}$	flour$_{F.1D}$/salt$_{F.3D}$	now
	b.	*isčez/isčezla*	*iz*	*magazinov.*	
		disappeared$_{M/F}$	from	stores	

[4] This study also involved aphasiac patients, while [Romanova and Gor 2017] compared native speakers to second language learners, but we will not discuss these groups here.

(2) a. *Nakanune* *(otëkšij)* *palec/lokot'* *sil'no…*
the-day-before swollen$_{M}$ finger$_{M.2D}$/elbow$_{M.2D}$ strongly
b. *bolel/bolela* *ot* *udara.*
hurt$_{M/F}$ from injury

[Slioussar and Malko 2016] studied gender agreement attraction. To give an example, an attraction error is present in the English sentence "The key to the cabinets are rusty", where the verb agrees not with the head of the subject phrase, but with another noun, termed *attractor*. In production, such errors are more frequent than agreement errors without attraction. In comprehension, they are missed more often and produce smaller delays in reading times and less pronounced ERP responses.

Number agreement attraction is widely discussed in the literature, while gender agreement has been analyzed only in a few studies so far. Among other things, it was noted that both in production and in comprehension, attraction effects can be observed in the sentences with singular heads and plural dependent nouns (e.g., "The key to the cabinets…"), but not in the sentences with plural heads and singular dependent nouns (e.g., "The keys to the cabinet…"). Almost all proposed explanations appeal to feature markedness, although approaches to markedness may be very different, from representational to frequency-based. Looking for similar asymmetries in gender agreement attraction, several studies of Romance languages obtained controversial results (e.g. [Acuña-Fariña et al., 2014]; [Anton-Mendez et al., 2002]; [Martin et al., 2014]; [Vigliocco & Franck, 1999]). [Badecker and Kuminiak 2007] found that neuter behaves as unmarked in a series of production experiments on Slovak, in which neuter is the least frequent gender, but is used in impersonal sentences, like in Russian.

[Slioussar and Malko 2016] conducted one production and three comprehension experiments. The results of the former were similar to the Slovak study, while in the latter, masculine behaved differently from feminine and neuter. Namely, attraction was observed for all dependent noun genders, but only for neuter and feminine heads. In other words, masculine heads were significantly more resistant to attraction: readers detected agreement errors irrespective of possible attractors' interference[5].

This result can be reconciled with the observations made in [Akhutina et al. 1999, 2001]; [Romanova & Gor 2017]. However, given that different patterns were observed for production and comprehension, we cannot explain them by a particular single property of gender features anymore. This reminds us that the notion of markedness usually invoked to explain all asymmetries between features is problematic because some studies rely on representational markedness (primarily counting the number of positive feature values in formal morphological models), the others consider the most frequent value to be the default etc. From the representational point of view, neuter is the unmarked gender in most accounts, while if we rely on frequency, masculine is. Maybe, these approaches should be seen as complementary, because different properties of features appear to be relevant in different experimental tasks.

[5] It is traditionally assumed that the features of the dependent noun are crucial for attraction, but both this study and some other findings suggest that the features of the head might be more important. We will not discuss this problem here.

Now let us turn to the problem of inflections that are more or less typical for a particular gender. Many experimental studies used different tasks (for example, determining a noun's gender or selecting an article) to show that nouns with more typical inflections are associated with faster and more accurate answers. This was demonstrated for Italian, French, Hebrew, Bulgarian etc. (e.g. [Andonova et al., 2004]; [Bates et al., 1995]; [Gollan & Frost, 2001]; [Spalek et al., 2008]).

However, fewer studies investigated the role of this factor in sentence processing. To give an example, Caffarra et al. (2015) looked at Italian nouns with more and less typical endings presented in the same sentences. They were preceded by articles, which carried gender information. Nouns from the two groups elicited different ERP responses. [Franck et al. 2008] and [Vigliocco and Zilli 1999] demonstrated for Italian, Spanish, and French that heads with regular inflections are more resistant to agreement attraction. The same is true for number agreement attraction (e.g.[Bock & Eberhard, 1993]; [Vigliocco et al., 1995]).

As for Russian, only [Taraban and Kempe's 1999] study discussed above addressed this problem. The authors found no differences between prototypical and non-prototypical subject nouns. However, the task they used required selecting a correctly agreeing verb form, which is a less immediate measure than simple reading times. For this reason, we decided to come back to this factor in the present study.

3. Experiment 1

This experiment was conducted together with Pavel Shilin, an MA student at Saint-Petersburg State University. The goal of the experiment was to find out whether subject–predicate gender agreement is processed differently depending on the declension and gender the subject noun belongs to. 33 native speakers of Russian aged 18–55 (12 women) took part in the experiment.

It is impossible to test all potentially interesting combinations of gender and declension in one experiment, so we selected the following three groups of inanimate nouns as subjects: masculine nouns of the 2nd declension (prototypical masculine, 2DM), feminine nouns of the 1st declension (prototypical feminine, 1DF), and feminine nouns of the 3rd declension (non-prototypical feminine, 3DF). The materials included 36 sets of target sentences in six conditions exemplified in (3a-f). All sentences had the same structure: a subject noun, *byl/byla* 'was$_{M/F}$', an adjective or participle, and a three-word PP.

(3)	a.	**2DM G(rammatical):**	*Xalat*	*byl*	*potrepannym ot mnogoletnej noski.*
			robe$_{M.NOM.SG}$	was$_{M}$	shabby$_{M}$ from years-long wear
	b.	**2DM U(ngrammatical):**	*Xalat*	*byla*	*potrepannoj…*
			robe$_{M.NOM.SG}$	was$_{F}$	shabby$_{F}$
	c.	**1DF G(rammatical):**	*Kurtka*	*byla*	*potrepannoj…*
			jacket$_{F.NOM.SG}$	was$_{F}$	shabby$_{F}$
	d.	**1DF U(ngrammatical):**	*Kurtka*	*byl*	*potrepannym…*
			jacket$_{F.NOM.SG}$	was$_{M}$	shabby$_{M}$

e. **3DF G G(rammatical):** *Šinel'* *byla* *potrepannoj…*
overcoat$_{F.NOM.SG}$ was$_{F}$ shabby$_{F}$
f. **3DF U(ngrammatical):** *Šinel'* *byl* *potrepannym…*
overcoat$_{F.NOM.SG}$ was$_{M}$ shabby$_{M}$
'The robe / jacket / overcoat was shabby from being worn for many years.'

Half of the sentences contained gender agreement errors on the predicate because taking previous studies of agreement into account (primarily agreement attraction experiments), the effects could be expected to be different in grammatical and ungrammatical sentences[6]. Subject nouns in the three declension groups were balanced for frequency and length using the *StimulStat* lexical database (http://stimul.cognitivestudies.ru, [Alexeeva et al., 2018]). Frequency information in this database is taken from the *Frequency Dictionary of Modern Russian Language* [Lyashevskaya & Sharoff, 2009].

Target sentences were distributed into six experimental lists so that each participant saw only one sentence from each set. The lists also contained 80 grammatically correct filler sentences. The sentences were presented on a PC using Presentation software (http://www.neurobs.com). We used the word-by-word self-paced reading methodology. Each trial began with a sentence in which all words were masked with dashes while spaces and punctuation marks remained intact. Participants were pressing the space bar to reveal a word and re-mask the previous one. One third of the sentences were followed by forced choice comprehension questions to ensure that the participants were reading properly.

We analyzed participants' question-answering accuracy and reading times. On average, participants answered 12% questions to target sentences incorrectly, no participants made more than 3 errors. Reading times that exceeded a threshold of 2.5 standard deviations, by region and condition, were excluded [Ratcliff, 1993]. In total, 2.0% of the data were excluded as outliers. Average reading times per region in different conditions are presented in **Figure 1**.

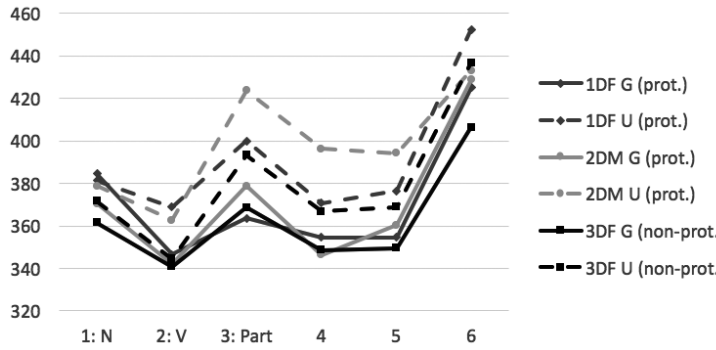


Figure 1. Average reading times per region (in ms) in different experimental conditions

[6] We selected predicates that consisted of a copula and an adjective or participle because such predicates were used in the previous experimental studies of subject—predicate gender agreement in Russian.

For each region, we made pairwise comparisons between the three conditions using a 2 × 2 Repeated Measures ANOVA with grammaticality and declension as factors. Analyses by participants (F_1) and by items (F_2) were performed. In region 1 (the subject noun), there were no significant results, which means that nouns in different conditions were properly balanced and that nouns of a particular gender or declension are not intrinsically more difficult to process.

Region 2 contains the verb *byl / byla* 'was$_{M/F}$'—this is where agreement errors appear in ungrammatical sentences. Figure 1 suggests that participants' reaction to these errors was different depending on the declension of the subject noun. In the conditions 1DF and 2DM (with prototypical feminine and masculine subjects), reading times in ungrammatical sentences are longer than in grammatical ones, while no such difference can be seen in the 3DF conditions (with non-prototypical feminine subjects), which indicates that the error tends to remain undetected in the latter case.

Statistical analyses support this intuition. In the comparison between 1DF and 3DF conditions, grammaticality and the interaction between declension and grammaticality are significant ($F_1(1,32) = 8.13$, $p < 0.01$, $F_2(1,35) = 4.20$, $p = 0.05$; $F_1(1,32) = 7.41$, $p = 0.01$, $F_2(1,35) = 4.05$, $p = 0.05$), while the main effect of declension does not reach significance. This means that the influence of grammaticality depends on the declension of the subject. In the comparison between 2DM and 3DF conditions, grammaticality reaches significance, while the interaction between declension and grammaticality is marginally significant ($F_1(1,32) = 8.01$, $p < 0.01$, $F_2(1,35) = 4.09$, $p = 0.05$; $F_1(1,32) = 3.98$, $p = 0.05$, $F_2(1,35) = 3.17$, $p = 0.08$). When 1DF and 2DM are compared, only the grammaticality factor is significant ($F_1(1,32) = 18.66$, $p < 0.01$, $F_2(1,35) = 10.21$, $p < 0.01$).

In region 3 that contains an adjective or participle, differences between grammatical and ungrammatical sentences become visible in all conditions. Accordingly, the grammaticality factor is significant in all pairwise comparisons ($F_1(1,32) = 15.90$, $p < 0.01$, $F_2(1,35) = 21.24$, $p < 0.01$ for 1DF vs. 2DM; $F_1(1,32) = 11.98$, $p < 0.01$, $F_2(1,35) = 6.20$, $p = 0.02$ for 1DF vs. 3DF; $F_1(1,32) = 9.73$, $p < 0.01$, $F_2(1,35) = 7.83$, $p < 0.01$ for 2DM vs. 3DF). No other factors or interactions reach significance.

Regions 4–6 contain a three-word PP. In region 4, a tendency that can be already detected in region 3 becomes statistically significant: the error-related delay in reading times is more pronounced in the 2DM conditions (with masculine subjects) than in the 1DF and 3DF conditions (with feminine subjects). In the comparison between 1DF and 2DM conditions, grammaticality and the interaction between declension and grammaticality are significant ($F_1(1,32) = 36.95$, $p < 0.01$, $F_2(1,35) = 15.91$, $p < 0.01$; $F_1(1,32) = 9.77$, $p < 0.01$, $F_2(1,35) = 6.45$, $p = 0.02$), while declension is not significant. The same is true for the comparison between 3DF and 2DM ($F_1(1,32) = 50.11$, $p < 0.01$, $F_2(1,35) = 13.17$, $p < 0.01$; $F_1(1,32) = 11.38$, $p < 0.01$, $F_2(1,35) = 5.51$, $p = 0.03$). When 1DF and 3DF are compared, only the grammaticality factor is marginally significant ($F_1(1,32) = 12.34$, $p < 0.01$, $F_2(1,35) = 3.65$, $p = 0.07$).

In region 5, only the grammaticality factor is significant in all pairwise comparisons ($F_1(1,32) = 18.51$, $p < 0.01$, $F_2(1,35) = 17.67$, $p < 0.01$ for 1DF vs. 2DM; $F_1(1,32) = 14.78$, $p < 0.01$, $F_2(1,35) = 6.10$, $p = 0.02$ for 1DF vs. 3DF; $F_1(1,32) = 18.07$, $p < 0.01$, $F_2(1,35) = 10.07$, $p < 0.01$ for 2DM vs. 3DF). In region 6, there are no significant differences.

Finally, let us note that when we planned the experiment, we did not consider assessing the role of the final consonant of 3DF nouns. But 10 out of 36 nouns we selected ended in *-l'* or *-r'*, which is more characteristic for masculine, while other nouns had final consonants characteristic for feminine. So the role of this factor could be estimated, and there were no hints of any relevant differences.

4. Pilot experiment 2

This pilot experiment was included in a study we conducted together with Varvara Magomedova (SUNY, Stony Brook) and Natalia Chuprasova, an MA student at Saint-Petersburg State University. The main goal of the study was to find out how Russian speakers determine the gender of real and nonce nouns with diminutive and augmentative suffixes. However, to make the materials more diverse, other nouns had to be included, and we selected 12 real and 12 nonce nouns ending in palatalized consonants (as well as some indeclinable nouns etc.).

Participants were 30 native Russian speakers (17 women), aged 19–30. They received a list of seven adjectives and then were presented with nouns one by one. They were asked to pick a matching adjective and pronounce the resulting phrase. Adjectives had meanings like 'big', 'small', 'cool', 'bad' etc., to make participants think that the experiment was about semantic connotations of different nouns.

Analyzing the gender of the adjectives selected by the participants, for 12 real nouns ending in palatalized consonants (6 masculine and 6 feminine) we found only 7 errors out of 360 responses. There were three errors with *žen'šen'* 'ginseng$_{M}$', two errors with *stupen'* 'step$_{F}$', and one error with *kisel'* 'starch drink$_{M}$' and with *prorub'* 'ice-hole$_{F}$'. The low number of errors agrees with the previous findings by Rusakova (2013): adult native speakers of Russian do not experience particular difficulties determining the gender of such nouns.

As for 12 nonce nouns, we had two examples with each of the following endings: *-b'*, *-d'*, *-s'*, *-n'*, *-l'* and *-r'*. The number of answers with masculine adjectives in these groups was 22 (out of 60), 26, 38, 30, 44 and 51, respectively. We can see that the nature of the final consonant played a role. To estimate it statistically, we used mixed effects logistic regression with random slopes and random intercepts by participants and by items. Only nonce nouns ending in *-l'* and *-r'* were significantly different from the other groups ($\beta = 2.48$, $SE = 1.15$, $z = -2.15$, $p = 0.03$; $\beta = 4.00$, $SE = 1.54$, $z = -2.60$, $p < 0.01$). Unlike all other palatalized consonants, these final consonants are more characteristic for masculine nouns.

5. Conclusions

In the introduction, we outlined three distributional properties of Russian nouns. Firstly, masculine gender is more frequent than feminine and neuter. Secondly, some combinations of genders and declensions are more frequent than the others (we called them prototypical). Thirdly, for nouns ending in palatalized consonants, some consonants are more characteristic for masculine nouns and the others for feminine nouns.

Previous studies indicate that these factors do not increase the number of naturally occurring gender agreement errors for adult native speakers of Russian [Rusakova 2013]. But speakers might still be sensitive to them, and in the present paper we demonstrated this in two experiments.

Experiment 1 revealed the role of the two first factors. It was not designed to assess the role of the third factor, but, as far as we could estimate from the data, this factor does not play a role in online agreement processing. However, the influence of this factor can be detected when speakers try to guess the gender of a novel noun—like in the pilot experiment 2. In further studies, we plan to use more stimuli, testing speakers' sensitivity to different suffixes etc.

Now let us discuss the results of experiment 1 in more detail. It demonstrated that both gender and declension of the noun influence online processing of the subject–predicate gender agreement in Russian. But, firstly, this influence can be detected only in the sentences with agreement errors, i.e. no gender or declension is intrinsically more difficult to process (at least, in the sentence context[7]). Secondly, declension plays a role at a very early stage and its effect is very short-lived, while the role of gender becomes visible later and its effect is more pronounced.

The fact that a masculine verb form is less readily detected after a 3rd declension subject noun can be explained by the fact that its ending is more typical for masculine nouns that for feminine ones. However, alternative explanations are also possible, for example, all agreement errors (in masculine or in neuter) may be harder to detect after 3rd declension subject nouns, i.e. their gender can be in general harder to retrieve. To exclude this and some other possibilities, other experiments should be conducted. Another line of further research should look at non-prototypical masculine nouns like *papa* 'dad'. The picture may be different not only because of their different gender, but also because all these nouns denote humans, so the gender feature is not semantically empty in this case, which may aid its processing and retrieval.

As for the role of gender as such, we saw that agreement errors with masculine subjects cause a larger delay in reading times compared to errors with feminine subjects, i.e. were costlier for processing. This is in line with the previous findings on gender agreement in comprehension reported in the literature [Akhutina et al., 1999, 2001]; [Romanova & Gor, 2017]; [Slioussar & Malko, 2016]. However, to have a full picture, neuter subjects and predicates should be introduced in further experiments.

So far, several explanations are possible. It is well known that while reading, we generate expectations about the upcoming predicate based on the features of the subject and rechecking is prompted if these expectations are violated (which is associated with increased reading times). Perhaps, the masculine form of the predicate, being the most frequent, causes less disruption if used incorrectly—similarly, using a frequent word instead of an infrequent one provokes less surprise than the opposite mistake. Maybe, these expectations are more robust for masculine subjects, so violating them is more disruptive. Maybe, if an agreement error is detected and rechecking is initiated, masculine subjects are retrieved more readily and reliably—this is what

[7] It is well known that many differences that can be detected in the processing of isolated forms disappear when these forms are embedded in an appropriate context.

[Slioussar and Malko 2016] suggested based on their agreement attraction results where all combinations of genders on subjects, attractors and predicates were examined. All these explanations are compatible with the observed difference between ungrammatical sentences with masculine and feminine subjects. Further experiments are necessary to tease them apart and to gain a better understanding of the patterns observed in previous studies.

The study was supported in part by the grant #16-18-02071 from the Russian Science Foundation.

References

1. *Acuña-Fariña, J. C., E. Meseguer, and M. Carreiras.* (2014). Gender and number agreement in comprehension in Spanish. Lingua 143: 108–128.
2. *Akhutina, Tatiana, Andrei Kurgansky, Maria Polinsky, and Elizabeth Bates.* (1999). Processing of grammatical gender in a three-gender system: Experimental evidence from Russian. Journal of Psycholinguistic Research 28: 695–713.
3. *Akhutina, Tatiana, Andrei Kurgansky, Marina Kurganskaya, Maria Polinsky, Natalya Polonskaya, Olga Larina, Elizabeth Bates, and Mark Appelbaum.* (2001). Processing of grammatical gender in normal and aphasic speakers of Russian. Cortex 37: 295–326.
4. *Alexeeva, Svetlana, Natalia Slioussar, and Daria Chernova.* (2018). StimulStat: a lexical database for Russian. To appear in: Behavior Research Methods.
5. *Andonova, E., S. D'Amico, A. Devescovi, and E. Bates.* (2004). Gender and lexical access in Bulgarian. Perception and Psychophysics 66: 496–507.
6. *Anton-Mendez, Inés, Janet Nicol, and Merrill F. Garrett.* (2002). The relation between gender and number agreement processing. Syntax 5: 1–25.
7. *Badecker, William and Frantisek Kuminiak.* (2007). Morphology, agreement and working memory retrieval in sentence production: Evidence from gender and case in Slovak. Journal of Memory and Language 56: 65–85.
8. *Bates, Elizabeth, Antonella Devescovi, Luigi Pizzamiglio, Simona D'Amico, and Arturo Hernandez.* (1995). Gender and lexical access in Italian. Perception and Psychophysics 57: 847–862.
9. *Bock, Kathryn, and Kathleen M. Eberhard.* (1993). Meaning, sound and syntax in English number agreement. Language and Cognitive Processes 8: 57–99.
10. *Caffarra, S., A. Siyanova-Chanturia, F. Pesciarelli, F. Vespignani, and C. Cacciari.* (2015). Is the noun ending a cue to grammatical gender processing? An ERP study on sentences in Italian. Psychophysiology 52: 1019–1030.
11. *Franck, Julie, Gabriella Vigliocco, Inés Antón-Méndez, Simona Collina, and Ulrich H. Frauenfelder.* (2008). The interplay of syntax and form in sentence production: a cross-linguistic study of form effects on agreement. Language and Cognitive Processes 23: 329–374.
12. *Gollan, T. H., and R. Frost.* (2001). Two routes to grammatical gender: Evidence from Hebrew. Journal of Psycholinguistic Research 30: 627–651.
13. *Janssen, Bibi E.* (2016). The acquisition of gender and case in Polish and Russian: A study of monolingual and bilingual children. Amsterdam: Pegasus.

14. *Lyashevskaya, Olga, and Sergey Sharov.* (2009). Častotnyj slovar' sovremennogo russkogo jazyka [The frequency dictionary of modern Russian language]. Moscow: Azbukovnik.
15. *Martin, A. E., M. S. Nieuwland, and M. Carreiras.* (2014). Agreement attraction during comprehension of grammatical sentences: ERP evidence from ellipsis. Brain and Language 135: 42–51.
16. *Ratcliff, Roger.* (1993). Methods for dealing with reaction time outliers. Psychological Bulletin 114: 510–532.
17. *Rodina, Yulia, and Marit Westergaard.* (2012). A cue-based approach to the acquisition of grammatical gender in Russian. Journal of Child Language 39: 1077–1106.
18. *Romanova, Natalia, and Kira Gor.* (2017). Processing of gender and number agreement in Russian as a second language. Studies in Second Language Acquisition 39: 97–128.
19. *Rusakova, Marina.* (2013). Elementy antropocentričeskoj grammatiki russkogo jazyka [Elements of the anthropocentric grammar of the Russian language]. Moscow: Jazyki slavjanskoj kul'tury.
20. *Savchuk, Svetlana.* (2011). Korpusnoe issledovanie variantov rodovoj prinadležnosti v russkom jazyke [A corpus study of morphological variability: variation in gender forms of Russian nouns]. Computer linguistics and intellectual technologies 10, 562–579.
21. *Schwartz, Mila, Miriam Minkov, Elena Dieser, Ekaterina Protassova, Victor Moin, and Maria Polinsky.* (2015). Acquisition of Russian gender agreement by monolingual and bilingual children. International Journal of Bilingualism 19: 726–752.
22. *Sekerina, Irina.* (2012). The effect of grammatical gender in Russian spoken-word recognition. In: Russian language studies in North America. New perspectives in theoretical and applied linguistics, ed. by Veronika Makarova, 107–132. New York: Anthem Press.
23. *Slioussar, Natalia, and Anton Malko.* (2016). Gender agreement attraction in Russian: production and comprehension evidence. Frontiers in Psychology 7: article 1651.
24. *Slioussar, Natalia, and Maria Samoilova.* (2015). Častotnosti različnyx grammatičeskix xarakteristik i okončanij u suščestvitel'nyx russkogo jazyka [Frequencies of different grammatical features and inflectional affixes in Russian nouns]. In: Proceedings of the conference 'Dialogue'. http://www.dialog-21.ru/digests/dialog2015/materials/pdf/SlioussarNASamoilovaMV.pdf
25. *Spalek, K., J. Franck, H. Schriefers, and U. H. Frauenfelder.* (2008). Phonological regularities and grammatical gender retrieval in spoken word recognition and word production. Journal of Psycholinguistic Research 37: 419–442.
26. *Taraban, Roman, and Vera Kempe.* (1999). Gender processing in native and non-native Russian speakers. Applied Psycholinguistics 20: 119–148.
27. *Tseitlin, Stella.* (2000). Jazyk i rebenok [Language and the child]. Moscow: Vlados.
28. *Vigliocco, Gabriella, and Julie Franck.* (1999). When sex and syntax go hand in hand: Gender agreement in language production. Journal of Memory and Language 40: 455–478.

29. *Vigliocco, Gabriella, and Tiziana Zilli.* (1999). Syntactic accuracy in sentence production: the case of gender disagreement in Italian language-impaired and unimpaired speakers. Journal of Psycholinguistic Research 28: 623–648.
30. *Vigliocco, Gabriella, Brian Butterworth and Carlo Semenza.* (1995). Constructing subject-verb agreement in speech: The role of semantic and morphological factors. Journal of Memory and Language 34: 186–215.
31. *Zaliznyak, Andrey.* (1987). Grammatičeskij slovar' russkogo jazyka. Slovoizmenenie [The grammatical dictionary of the Russian language. Inflection]. 2nd ed. Moscow: Russkij Jazyk.

IMPROVING NEURAL MORPHOLOGICAL TAGGING USING LANGUAGE MODELS[1]

Sorokin A. A. (alexey.sorokin@list.ru)
Moscow Institute of Physics and Technology, Dolgoprudnyj, Russia
Lomonosov Moscow State University, Moscow, Russia

We offer a new neural architecture for character-level morphological tagging, combining character-level networks with the output of neural language model on morhological tags. Our proposal reduces tagging error up to 10% in comparison with baseline model and achieves state-of-the-art performance both on ru_syntagrus and MorphoRuEval datasets.

Keywords: morphological analysis, tagging, neural network, neural language model, character-based models

АВТОМАТИЧЕСКИЙ МОРФОЛОГИЧЕСКИЙ АНАЛИЗ НА ОСНОВЕ НЕЙРОННЫХ МОДЕЛЕЙ С ИСПОЛЬЗОВАНИЕМ ЯЗЫКОВЫХ МОДЕЛЕЙ

Сорокин А. А. (alexey.sorokin@list.ru)
Московский физико-технический институт,
Долгопрудный, Россия
Московский государственный университет
им. М. В. Ломоносова, Москва, Россия

Данная работа посвящена автоматическому морфологическому анализу. Мы показываем, что комбинация символьных нейронных сетей с нейронными языковыми моделями улучшает качество морфологического анализа, снижая количество ошибок на 10%, при этом данный результат достигается без использования дополнительных ресурсов. Результат ещё улучшается в случае дополнительного использования морфологического словаря.

Ключевые слова: морфологический анализ, нейронные сети, символьные нейронные сети, нейронная языковая модель

[1] The research was conducted under support of National Technological Initiative Foundation and Sberbank of Russia. Project identifier 0000000007417F630002.

1. Introduction

There is no exaggeration in saying that last decade in computational linguistics is the decade of «neural network turn». The works of Tomas Mikolov, e.g. [Mikolov, 2013], on vector representations of words revolutionized not only computational semantics, but entire computational linguistics (further, CL), stimulating the fast growth of embedding-based approach. Another breakthrough insight was the introduction of character-based networks by Santos and Zadrozny [Santos and Zadrozny, 2014], permitting the researchers to solve practically every task from scratch provided enough data is available. After only several years, the vast majority of CL tasks of different complexity, from machine translation to morphological tagging, is solved mostly using different neural network-based architectures.

In the present paper we focus on the task of automatic morphological tagging which takes as input the sequence of words and assigns each word a label (or tag) containing the morphological description of that word. In early years of CL only the part-of-speech information was labeled, therefore this task is sometimes referred as POS-tagging. For analytical languages like English the sets for coarse part-of-speech tagging and fine-grained morphological labeling does not differ much in size and complexity. However, most of the languages are far more complex in its morphology and have a wider inventory of morphological categories, which makes the task of detailed morphological analysis much harder than coarse POS-tagging. Downstream applications benefit more from detailed morphological information (words connected by syntactic dependency often agree in their morphology, which cannot be revealed using only part-of-speech tags), therefore we address this very task and use the term morphological tagging through the paper.

Despite the undoubtable evidence for superior performance of neural network models for the plenty of tasks, their usage for morphological tagging worths further discussion. Indeed, most neural models were tested for English which has unbounded amount of training data and very simple morphology. For more complex morphology the patterns might be the opposite: for example, already the pioneer work [Lafferty et al., 2001] on conditional random fields was compatible with other morphological taggers. On the contrary, only clever design of learning process and output space made them capable to achieve state-of-the-art tagging level [Muller, 2013]. There is no a-priori evidence, that the same neural architectures are suitable for developed morphological structure of Russian or for small corpora in case of less widespread languages.

However, both the doubts are decisively disproved by recent research. What concerns the second problem, [Heigold et al., 2017] showed that a character-based neural tagger outperforms state-of-the-art CRF parser for a wide range of languages. The effect is rather clear even for training corpora with a thousand training sentences. That implies that LSTMs are more effective in capturing morphosyntactic patterns than CRFs possible due to their capability to learn long-distance dependencies.

The results of MorphoRuEval challenge [Sorokin et al., 2017] demonstrated that a deep neural model of [Anastasiev et al., 2017] defeats by a huge margin the second ranked tagger of [Sorokin and Yankovskaya, 2017], combining a hidden Markov model with linguistically motivated rules for reranking. Both these systems extensively used

external knowledge in the form of morphological dictionary, the first one also utilized the output of a closed rule-based semantic parser as feature, while the second heavily relied on feature engineering. Interestingly, other neural models except the winner were clearly behind second place.

Before describing our approach I would like to emphasize the difference between the behavior of neural and Markov tagging models on the example sentence

	его	*решение*	*задачи*	*было*	*неправильным*
(1)	*ego*	*reshenie*	*zadachi*	*bylo*	*nepravil'nym*
	his	solution+Sg+Masc	problem+Sg+Gen	be+Past+Sg+Neut	incorrect+Sg+Masc+Ins

Due to extensive regular homonymy in Russian it has more than 100 variants of tagging as summarized in the table below.

Table 1. Regular homonymy in Russian for the sentence *Его решение было неправильным*

Word	Number of tags	Tags
его	5	PRON, Gender=Masc, Case=Gen PRON, Gender=Masc, Case=Acc PRON, Gender=Neut, Case=Gen PRON, Gender=Neut, Case=Acc **DET**
решение	2	**NOUN, Gender=Neut, Case=Nom** NOUN, Gender=Neut, Case=Acc
задачи	3	**NOUN, Number=Sing, Case=Gen** NOUN, Number=Plur, Case=Nom NOUN, Number=Plur, Case=Acc
было	2	**AUX, Gender=Neut** PART
неправильным	3	ADJ, Number=Sing, Gender=Masc, Case=Ins **ADJ, Number=Sing, Gender=Neut, Case=Ins** ADJ, Number=Plur, Case=Dat

When parsing this sentence, an HMM relies on dictionary word-tag statistics and tag trigram frequencies. It decides that было should be an AUX since it is a dominant label of this word. An AUX,Gender = Neut tag is often followed by a neutral adjective in instrumental case and preceded by a NOUN, Gender=Neut, Case=Nom NOUN, Number=Sing, Case=Gen bigram. Finally, a determiner is more probable to occur before a noun, than a personal pronoun. Thus, the correct labeling is uncovered using only the tag cooccurence statistics. However, consider another sentence:

	его	*решение*	*задачи*	*будет*	*неправильным*
(2)	*ego*	*reshenie*	*zadachi*	*budet*	*nepravil'nym*
	his	solution+Sg+Masc	problem+Sg+Gen	be+Fut+Sg	incorrect+Sg+Masc+Ins

Russian verbs do not change by gender in non-past tense, consequently, the adjective *неправильным* has nothing to support the Gender=Neut hypothesis among its two preceding tags. Therefore HMM and CRF model are likely to fail since they do not take remote context into account.

On the contrary, neural models rely on the lexemes themselves, not the tag statistics. The word *было* (not its tag) forces the network to assign Number=Sing, Gender=Neut label to the next adjective *неправильным* and Case=Nom label to the word *решение* in its left vicinity. This latter word supports DET reading for *его* and *case=Gen* reading for *задачи* since genitives nouns often follow *решение*. Actually, this evidence is confirmed by other words ending by *–ние* as well since character-based networks capture graphical similatity. Moreover, a neural model probably captures the dependency between *решение* and *неправильным* making the second example less problematic.

Summarizing, lexical information is more important than grammar constraints that HMMs are trying to capture. The main advantage of neural network with respect to HMMs and CRFs is its ability to compress this information. It is usually "stored" in the states of a bidirectional LSTM, often being the principal layer of the model. To produce the probability distribution of word tags these states are usually passed through a one-layer perceptron with softmax activation. However, the tags predicted for different words do not directly affect each other. Consequently, the network has no direct mechanism to impose grammatical constraints on tag cooccurrences which may potentially limit its performance.

These constraints can in principle be learnt by neural language models on morphological tags. Such models are known to capture long-distance dependencies using memory mechanisms [Tran et al., 2016]. We propose two combinations which combine an underlying BiLSTM model of [Heigold et al., 2017] with a neural language model via the topmost layer of the tagger.

We apply our approach to UD2.0 [Nivre et al.] and MorphoRuEval [Sorokin et al., 2017] datasets for Russian language. Our paper is organized as follows: Section 2 introduces the baseline BiLSTM model, section 3 explains our extension of it, section 4 describes the experimental setup and presents tagging results, section 5 discusses the results obtained and we conclude with directions for future work.

2. Baseline model

Morphological tagging is a task of predicting a correct sequence of morphological tags $\boldsymbol{t} = t_1, ..., t_n$, given the words $\boldsymbol{v} = v_1, ..., v_n$. A character-based approach of [Heigold et al., 2017] addresses this problem from scratch and does not require any other resources except for a morphologically annotated corpus. Their model consists of two parts: the first encodes each word v_i (a sequence of characters) as a fixed-width embedding vector h_i, while the second transforms the obtained sequence of vectors $h_1, ..., h_n$ to the morphological tags.

First, we describe the encoding component of the model. The paper of Heigold uses two architectures for word representation: the first is a 2-layer LSTM while the

second combines several convolutional and highway layers. The first one slightly outperforms the second for most languages, however, we selected the second due to memory requirements. We refer the reader to the original paper for the full description, see also [Kim et al., 2016], applying the same ideas to neural language modelling. Briefly, the architecture is the following:

1. Each character is encoded as a 1-hot row vector with n_c dimensions, where n_c is the number of characters. A word of length L is represented by a sequence of L such vectors $x'_{i_1}, \ldots, x'_{i_L}$, that form a matrix X with L rows and n_c columns and exactly one unit in each row.
2. This matrix is multiplied by a matrix U of size $n_c \times n_e$, producing a sequence $X'=XU$ of L embeddings $x'_{i_1}, \ldots, x'_{i_L}$. j-th element of this sequence is a dense representation of i_j-th character in the alphabet.
3. X' is passed through K parallel convolutional layers with different widths. After this step K vectors of dimensions $f_1, \ldots, f_k$ are associated with each position of the word. Roughly speaking, k-th of these vectors contains information of useful ngrams of length w_k around current position.
4. All the vectors from the previous step are concatenated, producing a vector of length $F = \sum_j f_j$ for each symbol of the word. A word is now a matrix wit L rows and F columns.
5. A maximum-over-time (max-pooling) layer is applied to each row, finally encoding the word as a vector h' of fixed dimension F.
6. Several highway layers [Srivastava et al., 2015] are applied to this vector. Highway layer performs the transformation $h=s\odot(Vh')+(1-s)\odot h'$, where V is a square matrix with F rows, g is a non-linear function and $\odot$ denotes coordinate-wise product. The highway layer simultaneously produces useful combinations of features using one-layer perceptron (Vh') and keeps relevant dimensions of h'. The contribution of both components is balanced by means of vector s, which is obtained by another one-layer perceptron with sigmoid activation: $s=(Sh')$.

The second component of the network transforms the obtained sequence of word vectors $h_1, \ldots, h_n$ into n probability distributions $\pi_1, \ldots, \pi_n$. Here π_j contains tag probabilities for j-th word in the sentence. First, two LSTMs are applied, the first processing the sentence from left to right and the second from right to left. The first produces vectors $\vec{y}_1, \ldots, \vec{y}_n$ and the second outputs $\overleftarrow{y}_n, \ldots, \overleftarrow{y}_1$, thus each word is encoded by two vectors $\vec{y}_i, \overleftarrow{y}_i \in \mathbb{R}^{n_y}$. The concatenation of these vectors is multiplied by a projection matrix W with n_t rows and $2n_y$ columns, n_t being the number of tags. A softmax layer yields the required probability distribution:

$$y_i = [\vec{y}_i, \overleftarrow{y}_i] \quad \text{(concatenation)}$$

$$z_i = Wy_i$$

$$\pi_{ij} = \frac{e^{z_{ij}}}{\sum e^{z_{ik}}}$$

In [Heigold et al., 2017] this architecture is proved to be successful for languages of different morphological structure even with only several thousands of tagged sentences available for training.

3. Our proposal

3.1. Language models

As has already been said, the described architecture does not care about for the probability of the tag sequence "as a whole", it only tries to predict the most probable tag in each position. Hidden Markov models follow the opposite approach: they rewrite the probability of tag sequence given the word sequence as

$$p(t_1 \dots t_n \mid v_1 \dots v_n) \sim p(v_1 \dots v_n \mid t_1 \dots t_n) p(t_1 \dots t_n)$$

and further decompose this probability as

$$p(v_1 \dots v_n \mid t_1 \dots t_n) = p(v_1 \mid t_1) \dots p(v_n \mid t_n)$$

$$p(t_1 \dots t_n) = p(t_1) p(t_2 \mid t_1) p(t_3 \mid t_1 t_2) p(t_4 \mid t_2 t_3) \dots p(t_n \mid t_{n-2} t_{n-1})$$

assuming mutual independence of lexical probabilities $p(v_i|t_i)$. Morphological tags are supposed to be generated by a trigram model. Restricting lexical probabilities to single word-tag pairs is a drawback of HMMs since the morphological tag depends not only on the word it is assigned to, but also on the whole context (see discussion in the introduction). But proposed Char-LSTM architecture lacks this component at all, which limits its possibilities in an opposite way. Though, n-gram language models used in HMMs require much data for training and cannot access inner structure of morphological tags. However, language models can be based on neural networks as well, not only on n-grams: the probability of current tag t_i given the preceding tags $t_1,\dots,t_{i-1}$ might be obtained as an output of recursive neural network.

Neural language models were successful in modelling sequences of words (see [Tran et al., 2016], [Kim et al., 2016] and multiple references there) in large-scale tasks. We apply them to model sequences of morphological tags. We adapt the model of [Tran et al., 2016] which uses a variant of memory networks [Sukhbaatar et al., 2015] to attend the recent past.

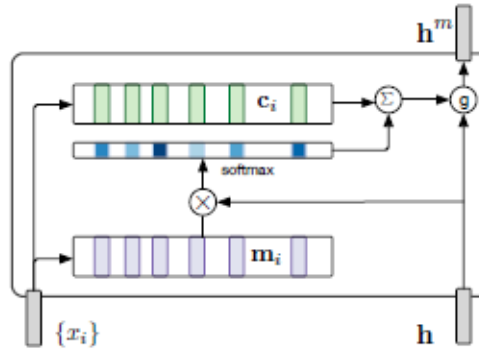


Figure 1. Memory block from [Tran et al., 2016]

The goal is to extract information which is the most relevant to predict further tags from the immediate left context of the current one. LSTM itself does this processing the sentence from left to right, but only partially, an additional memory block encoding context can capture more. The context in position i is a matrix X_i with d rows and m_v columns containing d preceeding elements $x_{i-d+1},\ldots,_i$. We multiply this context by two matrices M and C of size $m_v \times m_e$ obtaining two dense representations of the context $M_i=X_iM$ and $C_i=X_iC$. Actually, j-th row of M is the "input" embedding of j-th element in the vocabulary while j-th row of C is its output embedding. M_i is used to define the attention distribution over d preceding elements, which is calculated as:

$$\mathbf{p}_i = [p_{i,1}, \ldots, p_{i,d}] = \text{softmax}((M_i + T)h_i)$$

Informal explanation is the following: softmax favors those rows of $M_i=T$ which are the most similar to h_i. T is the bias which forces the model to attend particular positions independent from their content. Since h_i indirectly encodes the information about the past, the selected rows are the most relevant for this past. These rows should contribute the most to the context representation, so we use p_i as weights to produce an output representation of the context:

$$s_i = C_i^T p_i$$

s_i and h_i are concatenated to produce a joint embedding of the context in i-th position including both the global information from h_i and the relevant local information from s_i. As suggested in [Tran et al., 2016], this encoding is propagated through another LSTM layer:

$$\begin{aligned} h' &= [s_i, h_i], \\ z_i^{LM} &= LSTM(h'_1, \ldots, h'_i), \\ \pi_i^{LM} &= \text{softmax}(W_{LM} z_i^{LM}). \end{aligned}$$

In experiments of [Tran et al., 2016] x_i are just one-hot word encodings. However, morphological tags possess inner structure, therefore we apply encoding scheme summarized in Table 2. Additionally one can add an additional embedding layer before passing feature vectors to the LSTM. NOUN

Table 2. Input encoding of morphological tags

Feature dimension	Value
NOUN	1
VERB	0
…	0
NOUN, case=Nom	1
NOUN, case=Gen	0
ADJ, case=Nom	0
	0
NOUN, gender=Fem	1
NOUN, gender=Neut	0
NOUN, gender=Fem	0

Language model allows us to discriminate between probable and improbable combinations of tags, the next step is to apply it to the output of Char-LSTM to filter out inconsistent sequences.

3.2. Model combination

Now for each word in the sentence we have two probability distributions that predict its morphological tag. The first is the one of the Char-LSTM model, while the second generates current morphological label given already predicted tags $\pi^L(t_i)=p^{LM}(t_i|t_1 \ldots t_{i-1})$. They should be combined to produce the output distribution over tags. A naive way is to sum their logarithmic probabilities $\log(t_i)=\log \pi^{base}(t_i)+\log \pi^{LM}(t_i)$, assuming the independence of distributions under consideration. Obviously, these two distributions are not independent, therefore we take their weighted combination:

$$\log \pi(t_i) \sim s \log \pi^{base}(t_i) + (1-s) \log \pi^{LM}(t_i)$$

s itself is not a constant: obviously, the reliability of both distributions depends from internal states of corresponding models as well as from the position in the sentence. Informally, for some words the Char-LSTM model is already a good predictor, so the language model weight should not be large. In the beginning of the sentence neural LM is also irrelevant since there is no history it can rely on. On the contrary, when observing rare or homonymous words we should trust the LM more. Summarizing, we choose s to be a vector of weights, not a single weight. It is predicted using a single-layer perceptron with sigmoid activation:

$$\begin{aligned} z_i^w &= [z_i^{base}, z_i^{LM}, pos_i], \\ \mathrm{s}_i &= \sigma(S^w z_i^w + b^w) \end{aligned}$$

Here $pos_i=log(1+i)$ is a scalar encoding current position; z_i^{base} and z_i^{LM} are the states of the topmost LSTMs for Char-LSTM tagger and neural LM, respectively; s_i and b^w are vectors of dimension n_{tags} and S^w is a matrix with n_{tags} rows and $d^{base}+d^{LM}+1$ columns. Here d^{base} and d^{LM} are hidden state dimensions for char-LSTM and neural LM,

respectively. In principle, a multilayer network instead of a single layer could be applied. We refer to this architecture as Char-Weight in the further.

However, the weighting scheme helps only if at least one probability distribution is reliable, it is not capable to correct synchronous errors. Analogously to [Gulcehre, 2015], we use another approach. The output distributions of both the neural LM and the CharLSTM tagger are obtained by projecting their states by means of one-layer perceptron. It implies that all the probability information is already encoded in these states. The idea is to fuse the states of CharLSTM and neural LM into a single state and then process it using a separate network. We choose a two-layer perceptron with ReLU activation as such a network, formally:

$$\begin{aligned} z_i^w &= [z_i^{base}, z_i^{LM}, pos_i], \\ z'^w_i &= \max(S_1 z_i^w + b_1, 0), \\ \pi_i &= \mathrm{softmax}(S_2 z'^w_i + b_2). \end{aligned}$$

We refer to the second model as CharFusion.

4. Experiments and Results

4.1. Experimental setup

For CharLSTM model we use the setup of [Heigold et al., 2017] with minor modifications. Namely, character encoding dimension is 32, there are 7 convolutional layers applied in parallel with their width ranging from 1 to 7. The number of filters on layer with width w is min(200, 500w), so each position of the word is encoded by vector with 1100 elements after passing the convolution. On the word level we use LSTMs with 128 units in each direction. To prevent overfitting we apply dropout to word embeddings and to the outputs of topmost LSTM layer, the dropout probability is 0.2. We use the shallow variant of the architecture, which means only 1 convolutional and highway layers are applied on character level and only one LSTM layer on the word level.

In the neural language model dense tag embeddings have dimension 96 as well as the memory embeddings in the attention layer, the history window to be attended is 5. Output LSTM has 128 hidden units. 0.2 dropout is applied to the outputs of all embedding layers. In the CharFusion model we use 256 units on the hidden layer of the output perceptron.

All models are implemented in Keras library [Chollet et al.] with Tensorflow backend. The models are optimized using Adam optimizer with Nesterov momentum [Dozat, 2016], the learning rate and other optimizer parameters are set to default. The taggers are trained for 75 epochs, language models are trained for 50 epochs, the conventional cross-entropy loss (negative logarithmic probability of correct sequence) is used. When training CharWeight and CharFusion models, we train the basic CharLSTM component of them as well, the weight of the basic model loss is 0.25. We stop

training when the loss on development set have not improve for 10 epochs, saving the model with the best performance on the validation set.

We did not perform exhaustive hyperparameter search. However, preliminary experiments has shown that character embeddings of size 16 as in the original paper lead to worse performance and using recurrent networks with more hidden states or layers slightly deteriorates tagging accuracy. We have also found that regularizing the output probability distribution with L2 loss makes this distribution smoother and prevents overfitting, the regularization coefficient was set to 0.005.

Searching for the optimal tag in position i requires the knowledge of preceding morphological label. In the training phase we feed the model with golden tags, which are not available in test time. Therefore during testing we predict the tags one-by-one from left-to-right and return the sequence with maximal sum of logarithmic tag probabilities. To make the model capable to recover from its errors we apply a beam search with beam width 5. That raises the problem of exposure bias: when training, the model sees only the correct tags as the left context. However, if in the test phase the models fails to predict a correct tag in position i, all the predictions in positions i+1, i+2, ... will be done with incorrect tag history. Neither the tagger, nor the language model, are able to deal with such histories since they were trained only on gold contexts. This problem is called the exposure bias, to alleviate it we replace a 20% fraction of tags in the left context by a vector of all zeros forcing the model to operate correctly even if it lacks complete information about tag history.

4.2. Dataset

We evaluate our model on ru_syntagrus subcorpus of Universal Dependencies 2.0 corpus [Nivre et al.], the train subsection was used for training, the development one for validation and the test part for evaluation. We lowercase all the words, in case a word starts with a capital letter or consists of all capitals special pseudoletters <FIRST_UPPER> or <ALL_UPPER> were added in the beginning. All the letters appearing less than 3 times were replaced by special <UNK> symbol.

The size of the corpus in sentences and words is given in 3. Experimental results are presented in Table 4, we evaluate both per-tag and per-sentence accuracy. We observe that CharWeight model reduces error rate by about 6% depending on the corpus while the CharFusion error reduction exceeds 10%. It demonstrates that our model indeed improves the quality of morphological tagging.

Table 3. ru_syntagrus corpus statistics

Corpus	Words	Sentences
Train	870,033	48,814
Development	118,427	6,584
Test	117,276	6,491

Table 4. Evaluation on UD2.0 dataset. ERR—error rate reduction

Model	Tag accuracy	ERR	Sentence accuracy	ERR
CharLSTM (baseline)	95.19	0.0	52.22	0.0
	95.22	0.0	50.98	0.0
CharWeight	95.54	7.3	54.95	5.7
	95.52	6.3	53.66	5.5
CharFusion	95.70	10.6	57.15	10.3
	95.70	10.0	56.29	10.8

4.3. Using morphological dictionary

Roughly speaking, morphological tagging for dictionary words simply selects the most appropriate tag from a predefined set of dictionary tags of the current word. Therefore we enrich the data which the model accesses with the output of morphological analyzer PyMorphy [Korobov, 2015]. For each word we compute the set of possible tags using PyMorphy, transform these tags to UD2.0 format by means of freely available russian-tagsets package and extract all UD2.0 categories that are compatible with the labels obtained. The list of categories is encoded using one-hot scheme and then embedded into a dense vector of length 256. This vector is concatenated to word embedding that is obtained from the character-level network.

Table 5 contains results of model evaluation in case a morphological dictionary is added. We find that there is no clear gain from using a language model in this case. This effect is surprising to us and we plan to investigate it further.

Table 5. Tagging accuracy when using a language model

Model	Tag accuracy	ERR	Sentence accuracy	ERR
CharLSTM(baseline)	95.19	0.0	52.22	0.0
	95.22	0.0	50.98	0.0
CharLSTM+PyMorphy	96.30	23.0	60.48	17.3
	96.43	25.3	60.01	18.4
CharWeight+PyMorphy	96.26	22.2	60.65	17.6
	96.43	25.3	60.21	18.8
CharFusion+PyMorphy	96.34	23.9	61.80	20.0
	96.46	25.8	60.70	19.8

5. MorphoRuEval 2017 Dataset

We also evaluate our models on MorphoRuEval-2017 dataset [Sorokin et al., 2017]. We compare against two best models, the deep learning one of [Anastasiev et al., 2017] and the HMM-based rule reranker [Sorokin and Yankovskaya, 2017]. The results of comparison are in Table 6. We used the results mentioned in the papers and the official evaluation script of the contest.

Table 6. Results on MorphoRuEval-2017 dataset

Model	MorphoRuEval dev		MorphoRuEval test	
	Tags	Sentences	Tags	Sentences
Anastasiev et al., 2017	97.8	NA	97.1	83.3
Sorokin, Yankovskaya, 2017	96.3	78.5	94.8	69.3
CharLSTM [Heigold et al., 2017]	95.8	74.9	94.6	67.0
CharFusion	96.1	77.0	94.9	68.0
CharLSTM+PyMorphy	96.3	77.4	95.1	68.8
CharFusion+PyMorphy	96.6	79.8	95.4	71.1

We observe that our best model outperforms the second system of MorphoRuEval-2017 being sufficiently behind the first one. Note, that the model of [Anastasiev et al., 2017] used an additional training corpus and complex representations from in-house parser. Therefore our tagger demonstrates state-of-the-art performance on MorphoRuEval dataset as well.

6. Conclusions and future work

The present work mainly is a "proof-of-concept": we have demonstrated that character-level morphological tagging can be significantly improved using neural language models on morphological tags. Our work establishes a new state-of-the-art for tagging from scratch without access to external morphological resources. The natural direction is to test our approach on other languages with less data available analogously to the previous work. However, tagging almost a half of the sentences erroneously is a significant problem. Actually, some of these errors are not relevant since UD morphological tags contain categories which cannot be determined from the context (e. g., verb aspect) or does not have a clear bound from other categories (e. g., participles, which are treated as verbs), therefore it would be more natural to exclude them from being evaluated.

We have also demonstrated that adding the information from morphological dictionary can further improve performance. Actually, we have tried the simplest way to do it and further analysis is required. Another way to boost the model is to utilize task-independent embeddings obtained from large unlabeled corpora. The next challenge is to achieve the state-of-the art quality of closed systems using only open resources. We plan to address this question in the future work.

Another direction of research is improving neural models for morphological tags. Actually, not all government constraints can be addressed by the language model. For example, prepositions in Russian require different cases to their right ("*без друга*" *without the friend*+Gen vs "*про друга*» *about the friend*+Acc). The information about preposition cases is not encoded in their UD tags therefore in this case the CharLSTM component has to do the job more appropriate to the tag language model. Even a harder problem arises with verb government, consider *солгал другу* vs *обманул друга* both meaning "told+3 a lie to a friend" but with different case forms of the word *друг* ("*friend*"). Such examples demonstrate that actually we cannot separate "lexical"

and "morphological" part of tagging models. Probably, morphological tagging should be tackled using more complex architectures for sequence-to-sequence learning.

Summarizing, we have introduced a straightforward and linguistically motivated way to improve the quality of morphological tagging without having access to any external resources except the annotated corpus. Using external morphological dictionaries further improves performance. Our architecture is language-independent and does not have task-specific parameters, which makes it useful for applying "out of the box".

7. Acknowledgements

The author thanks Ekaterina Yankovskaya for invaluable help in editing and improving the first version of the paper. He is also grateful to his collegues in the laboratory of Neural Systems and Deep Learning of Moscow Institute of Physics and Technology and especially to Mikhail Arkhipov for helpful discussions.

References

1. *Anastasiev D. G., Andrianov A. I., Indenbom E. M.* (2017), Part-of-speech tagging with rich language description, Computational linguistics and intellectual technologies: Proceedings of the International Conference "Dialog 2017". [Komp'yuternaya lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2017"], Moskva, pp. 2–13. http://www.dialog-21.ru/media/3895/anastasyevdgetal.pdf
2. *Gulcehre C. et al.* (2015), On using monolingual corpora in neural machine translation. arXiv preprint arXiv:1503.03535.
3. *Dozat T.* (2016), Incorporating nesterov momentum into adam. Available at https://openreview.net/pdf?id=OM0jvwB8jIp57ZJjtNEZ
4. *Heigold G., Neumann G., van Genabith J.* (2017), An extensive empirical evaluation of character-based morphological tagging for 14 languages, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Long Papers, Vol. 1., pp. 505–513.
5. *Chollet F. et al.,* Keras, available at https://github.com/keras-team/keras
6. *Kim Y., Jernite Y., Sontag D., Rush A. M.* (2016), Character-Aware Neural Language Models, AAAI, pp. 2741–2749.
7. *Lafferty J., McCallum A., Pereira F. C. N.* (2001), Conditional random fields: Probabilistic models for segmenting and labeling sequence data, available at https://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis_papers
8. *Mikolov T., Chen K., Korrado G., Dean J.* (2013), Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.
9. *Müller T., Schmid H., Schütze H.* (2013), Efficient higher-order CRFs for morphological tagging, Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, pp. 322–332.
10. *Nivre J. et al.,* (2017), Universal dependencies 2.0., available at https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1983

11. *Santos C. D., Zadrozny B.* Learning character-level representations for part-of-speech tagging (2014), Proceedings of the 31st International Conference on Machine Learning (ICML-14), Bejiing, pp. 1818–1826.
12. *Sorokin A. A. et al.* (2017), MorphoRuEval-2017: an evaluation track for the automatic morphological analysis methods for Russian, Computational linguistics and intellectual technologies: Proceedings of the International Conference "Dialog 2017". [Komp'yuternaya lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2017"], Moscow, pp. 297–313, available at http://www.dialog-21.ru/media/3951/sorokinaetal.pdf
13. *Sorokin A. A., Yankovskaya E. V.* (2017), Using Context Features for Morphological Analysis of Russian, available at https://www.researchgate.net/publication/319623361_Using_Context_Features_for_Morphological_Analysis_of_Russian
14. *Srivastava R. K., Greff K., Schmidhuber J.* (2015), Highway networks, arXiv preprint arXiv:1505.00387.
15. *Sukhbaatar S., Szlam A., Weston J., Fergus R.* (2015), End-to-end memory networks, Advances in neural information processing systems, Montreal, pp. 2440–2448.
16. *Tran K., Bisazza A., Monz C.* (2016), Recurrent memory networks for language modeling, arXiv preprint arXiv:1601.01272.

DIFFERENTIAL OBJECT MARKING IN CONTACT-INFLUENCED RUSSIAN SPEECH: EVIDENCE FROM THE CORPUS OF CONTACT-INFLUENCED RUSSIAN SPEECH OF RUSSIAN FAR EAST AND NORTHERN SIBERIA[1,2]

Stoynova N. M. (stoynova@yandex.ru)
Vinogradov Russian Language Institute, RAS; Moscow, Russia

The paper deals with differential object marking in the Russian Speech of Nanai-Russian bilingual speakers, namely the variation such as *принес рыбу ~ принес рыба* ('{he} brought fish-acc ~ fish-nom'). The puzzle is that this peculiarity can result from a number of different processes: morpho-syntactic borrowing from Nanai, penetration of dialectal features into the speech of bilinguals, under-acquisition or reinterpretation of the Standard Russian system. The data of a small corpus of contact-influenced Russian Speech is used to test all these hypotheses. The results are following. Nominative forms are used in DO-position in quite a systematic way and such uses cannot be estimated as occasional "errors". The main factors that influence the NOM~ACC distribution are a) information structure and b) the accentual type of noun stem. The latter fact supports the hypothesis of a systematic reinterpetation of the Standard Russian system in the situation of incomplete acquisition. No significant correlations with animacy, definiteness, verb form and word order were attested. DOM pattern of Nanai Russian differs from those of Russian dialects and reveals some similarity to those of Nanai. However it cannot be considered as a full morphosyntactic calque.

Keywords: Russian, differential object marking, corpus linguistics, language contact

[1] The research was conducted with support of RSF grant No. 17-18-01649 (Dynamics of language contact in the circumpolar region).

[2] Many thanks to my colleagues S. Oskolskaya, I. Khomchenkova, P. Pleshak and A. Shluinsky, to my Nanai-speaking consultants, especially to VSG and to the anonymous reviewers of "Dialogue-2018".

ДИФФЕРЕНЦИРОВАННОЕ МАРКИРОВАНИЕ ОБЪЕКТА В КОНТАКТНО-ОБУСЛОВЛЕННОЙ РУССКОЙ РЕЧИ: ДАННЫЕ КОРПУСА КОНТАКТНО-ОБУСЛОВЛЕННОЙ РУССКОЙ РЕЧИ СЕВЕРА СИБИРИ И ДАЛЬНЕГО ВОСТОКА

Стойнова Н. М. (stoynova@yandex.ru)
Институт русского языка им. В. В. Виноградова, РАН;
Москва, Россия

В докладе рассматриваются случаи дифференцированного маркирования объекта, характерные для русской речи русско-нанайских билингвов, ср. конкурирующие структуры вроде *принес рыбу* ~ *принес рыба*. Эти случаи интересны тем, что не до конца понятно, чем они мотивированы: непосредственным влиянием первого языка (нанайского), неполным усвоением русского или явлениями неконтактной природы — диалектными особенностями локальной разновидности русского языка. Для исследования этого вопроса привлекаются данные создаваемого нами небольшого корпуса контактно-обусловленной русской речи. Эти данные обнаруживают следующую картину. Система DOM в нанайском русском кажется достаточно последовательной. Основными факторами, регулирующими выбор между номинативом и аккузативом в прямообъектной позиции оказываются информационная структура и акцентный тип основы. Последнее можно считать аргументом в пользу гипотезы о системной реинтерпретации русской системы в условиях неполного усвоения языка. Не обнаружено значимых корреляций с одушевленностью, определенностью, порядком слов, формой вершинного предиката. Система DOM в нанайском русском заметно отличается от представленной в русских диалектах (что не позволяет принять гипотезу о диалектном субстрате) и обнаруживает сходство с системой, представленной в нанайском языке (однако оно не достаточно для того, чтобы считать этот случай чистым случаем прямого морфосинтаксического калькирования).

Ключевые слова: русский язык, корпусная лингвистика, языковые контакты, грамматическая интерференция, дифференцированное маркирование объекта

0. Introduction

The paper has two main goals. The first goal is to present an ongoing project of creating a corpus of Contact-influenced Russian Speech of Russian Far East and Northern Siberia. The second goal is to show how the data of the corpus can be used

in order to draw the borderline between a true grammatical interference and peculiarities of other origin attested in the Russian Speech of bilinguals.

The corpus of contact-influenced Russian Speech contains by the moment ca. 20 hours of oral speech (mostly narratives) of bilingual speakers of Samoyedic and Tungusic languages. The texts are transcribed in standard Russian orthography and supplied with a morphological annotation and a manual annotation of grammatical peculiarities.

The case study presented in the paper deals with differential object marking (DOM) attested in texts of the corpus. Some of bilingual speakers (both speakers of Samoyedic and Tungusic) widely use nominative in Direct Object position, as in (1), as well as the expected accusative, as in (2):[3]

(1) *Рыба* сдаем/ (NOM)

(2) То плохую *рыбу*\ принесла — чо\ там, собакам\ буду варить (ACC)

At first glance, 1) Nominative and Accusative are used as free variants with no strict distribution, 2) the predisposition to Nominative forms in DO-position varies greatly across languages in contact and across individual speakers, so the general picture seems to be chaotic.

This is why an investigation of this feature should be a) held based on a text corpus (as large as possible by the moment), b) started by an analysis of particular idiolectal sub-systems.

In this paper I present an analysis of detailed data recorded from only one speaker. This speaker (VSG, 1931, the village of Kharpichan, Khabarovsk Krai) is fluent both in Nanai (Southern Tungusic) and in Russian; she learnt Russian at school (3 years) and now uses mostly Russian in her everyday life. Her Russian Speech reveals lots of deviations from Standard Russian which are presumably contact-induced. I analyze the full sample of DOs attested in her speech (see 2.3 below). The term "Nanai Russian Speech" is used in this paper for the Russian Speech of VSG.

This approach encompasses some general problems of extremely small fieldwork text collections (cf. Ostler 2008; Cox 2011; Mosel 2014; Vinogradov 2016 among others). Such collections are not as well-balanced and representative as standard large text corpora. However this is often the only type of text data available. There are two main risks in the case under discussion. First, the results of a study on one-speaker-corpus cannot be extrapolated with confidence to all patterns of speech of Nanai-Russian bilinguals. Second, in such a small text sample the quantitative analysis may be biased by particular genres, particular texts and particular lexical items, used in these texts. In this study I will not test the reliability of my data. However, having taken into account these risks. I we will try to estimate if such data can give any plausible results.

The nature of DOM in Nanai Russian Speech is not self-evident. The following hypotheses can be proposed.

---

[3] Similar patterns of DOM are attested also in other contact-influenced varieties of Russian, cf. Daniel et al. 2010: 81 on Daghestanian Russian. However it is not evident that all such cases are in fact of the same nature. E.g. in this paper the pattern is analyzed as "quasi-ergativity" (the result of the interference with the ergative alignment of Nakh-Daghestanian L1's).

1) It may be a direct morphosyntactic calque (pattern-borrowing) from Nanai.
2) It may be a result of under-acquisition of the Russian case system by bilingual speakers with no clear prototype in their L1.
3) It may be not of a contact nature at all: similar syntactic patterns are attested in non-contact Russian dialectal varieties.

Finally, all these potential sources of DOM may play a role and interact with one another.

1. Preliminary remarks on DOM

1.1. Cross-linguistic expectations for DOM

Differential object marking (DOM) is a situation in which a direct object (DO) can be marked with two or more competing forms. It is very widespread across languages of the world and well-studied in a cross-linguistic perspective (cf. [Bossong 1985]; [Aissen 2003]; [Malchukov, de Swart 2009]; [Witzlack-Makarevich, Seržant 2017] among many others).

The choice between competing forms can be strict or not (split DOM vs. fluid DOM, [Malchukov, de Swart 2009]). Competing forms can be both equally marked (symmetric DOM), however the case where one of them is unmarked (asymmetric DOM) is quite typical. The choice can be regulated by inherent or contextual properties of the direct object itself vs. by features of the predicate (argument-triggered DOM vs. predicate-triggered DOM, [Witzlack-Makarevich, Seržant 2017]). In particular, the following factors can be relevant.

a) Inherent semantic features of the direct object, such as human vs. non-human, animacy, uniqueness, discreteness; splits are expected to follow the Silverstein's hierarchy or similar hierarchies:

(3) personal pronouns > proper names > humans > animals > inanimate objects [Silverstein 1976]

b) Definiteness and specificity (referential properties) of direct the object; splits are expected to follow the hierarchy of definiteness:

(4) definite objects > specific indefinite objects > non-specific indefinite objects

c) Information structure (cf. [Dalrymple, Nikolaeva 2011]; [Iemmolo 2010] for the discussion).
d) Such features of the head predicate as finiteness, TAM, polarity and others.

The theoretical discussion on DOM focuses mainly on its possible functional motivations. DOM is considered either as a way to signal the semantic features of direct object themselves (indexing function of DOM) or as a way to disambiguate between the direct object and the subject within the clause (differentiating function of DOM), cf. e.g. [Malchukov 2008].

An attempt to involve the data of bilingual speech can bring a new dimension to the discussion. In this case additional motivations to follow the pattern of L1 or to re-analyze the system of L2 is added.

1.2. DOM in Nanai

DOM is attested in Nanai (the first language of the speaker under discussion, VSG). Dedicated accusative forms with the marker *-wA~-bA* compete with nominative (unmarked) forms in DO-position. The choice is not strict (fluid DOM). The following factors are relevant for the choice between nominative vs. accusative marking of DO.

1) Definiteness and specificity: NOM is more frequent for indefinite specific objects and especially for indefinite non-specific ones.
2) Information structure: NOM is less frequent with the topic marker *=tAni*.
3) Number: NOM is less frequent for objects with the plural marker (presumably due to formal rather than semantic reasons).
4) Phonetic context: NOM is more frequent in the context of words on *wA-* (which are phonetically similar to the accusative affix).

For more detail see Avrorin 1948: 223–233 and Oskolskaya, Stoynova 2017.

1.3. DOM in Russian dialects

One more potential source of non-standard marking of direct object in Nanai Russian is dialectal substrate. So called "nominative object constructions", as in (5), are attested in some Russian dialects, cf. Markova 1989; Ron'ko 2017 among others.

(5) Вам только гроб сделать да *яма выкопать*.
[Свадьба (Архангельская область, 1994), RNC[4]]

As [Ron'ko 2017] points out,

1) This feature is attested in different dialectal groups; it is especially characteristic for Northern dialects, however not only for them.
2) The main context for nominative objects are infinitive constructions (such as in (5)), to a lesser extent they are also used in finite clauses.
3) The choice between NOM and ACC is free, however NOM is more frequent (at least in Northern dialects):
    a) for indefinite objects and especially for non-specific ones;
    b) for foci, rather than for topics;
    c) for objects of clauses with OV word order.

A reason to suspect a non-contact nature of DOM in the Russian Speech of VSG (and in Nanai Russian in general) is the presence of other dialectal or regional features in her Russian Speech. Currently I have not enough data to attribute these features with confidence to a specific dialect group, they can be of a mixed nature. See examples of: a) lexical dialectal features: *мамка*, *папка*, *маленько*; b) phonetic ones:

[4] www.ruscorpora.ru.

[o] in unstressed syllables[5]; c) morphosyntactic ones: the genitive forms of pronouns *мене*, *тебе* (instead of *меня*, *тебя*), the preposition *с* instead of *из* (*с Москвы*).

VSG was born and spent her childhood in the village of Kondon (the settlement of Sorgolj), Solnechnyj District. There she acquired Russian at school-age. I have no clear data on the Russian input of VSG (1931) and her age-mates from the area. In the 1930-s, the village was inhabited almost only by Nanai people. Contacts with Russians were not intensive in the first half of 20th century. The closest large Russian village was Nizhnjaja Tambovka, inhabited since the middle of 19th century by colonists from Tambov province. The first Russian teachers of Kondon's school (opened in 1902) were from this village. In the 1930s a group of Russian Communist party activists (so called "Krasnaja Jurta") worked in Kondon. The active invasion of Russians from all over the country into the area near Kondon began in the end of the 1930s. In 1938, the construction of Baikal-Amur Mainline started not far from Kondon. At the same time, in 5–10 kilometers from Kondon a subdivision of the Gulag camp ("NizhAmurLag") was settled (nowadays, the village of Kharpichan where VSG lives now).

Given this background, an intensive influence of Northern Russian dialects is hardly probable. Still, traces of other dialects (or of a mixture of dialects) are possible.

Below I address the data on Russian dialects as comparative well-studied data on a similar DOM strategy rather than as a possible source of the pattern under discussion.

2. The system of DOM in Nanai Russian

2.1. Nominative in DO-position among other non-standard uses of Nominative

The use in DO-position is not the only (though the most frequent) non-standard use of Nominative attested in Nanai Russian Speech. Nominative is also used in the text sample:

a) in a numeral phrase;
b) in a possessive construction for the possessor;
c) for noun attributes;
d) in a preposition phrase (mostly with prepositions that take Genitive in Standard Russian);
e) in the negative existential construction;
f) rarely, in a verbal argument position where cases other than Accusative or PPs are used in Standard Russian. Presumably these uses are driven by a non-standard information structure.

In the majority of the listed non-standard uses Nominative corresponds to Genitive in Standard Russian and to Nominative (or rather to the unmarked form) in Nanai.

[5] It can be however not dialectal "okanje", but phonetic interference with Nanai which has no such type of vowel reduction, as Standard Russian.

Table 1 shows the frequencies of different types of uses, the list of correspondences in Standard Russian and in Nanai and examples.

Table 1. Different non-standard uses of Nominative in Nanai Russian Speech

	% (N)	in Standard Russian	in Nanai	example
DO	48% (39)	acc	acc~nom	*рыба сдаем*
num phrase	16% (13)	gen	nom	*три доска*
PP	10% (8)	*без*, *от*, *мимо*, *после* + gen, *по* + dat	nom	*мимо бабушка\ иду*
possessor	7% (6)	gen	nom	*кета шкурой*
exneg	7% (6)	gen	nom	*краснота нету*
attribute	2% (2)	different	nom	*апрель месяце*
other uses, presumably motivated by information structure	9% (7)	different	different	*Крапива мы огород\ удобряем Этот март рак\ она умерла*
total	100% (81)			

2.2. DOM in Nanai Russian compared to the Standard Russian system

Animacy distinction expressed within the case system of Standard Russian can be also interpreted as a case of DOM, but it differs significantly from what is observed in Nanai Russian Speech. Unlike Nanai Russian, in Modern Standard Russian:

1) there is no free variation, but there is a strict split, conditioned by animacy only (ACC=NOM for inanimate nouns, ACC=GEN for animate nouns).
2) The split is relevant only for a part of noun paradigm, namely for plural stems and for masculine singular zero-stems.
3) This is a morphological split in case marking rather than a syntactic one: the accusative form is equal to NOM~GEN not only in DO-position, but also in other contexts typical of the accusative case in Russian, including prepositional phrases.

Table 2. DO-marking: Nanai Russian vs. Standard Russian

	Standard Russian	Nanai Russian
inanimate pl	=NOM (*вижу столы*)	NOM
animate pl	=GEN (*вижу слонов*)	GEN~NOM
inanimate 0-stems sg	=NOM (*вижу стол*)	NOM
animate 0-stems sg	=GEN (*вижу слона*)	GEN~NOM
o-stems sg	=NOM (*вижу окно / чудовище*)	NOM
0-stems feminine sg	=NOM (*вижу мать / печь*)	NOM
a-stems sg	ACC (*вижу маму / печку*)	ACC~NOM

As shown in Table 2, a) the Nanai Russian data of our sample form quite a consistent system, b) this system retains the distinction attested in Standard Russian. It can be described as the Standard Russian system complicated with an additional option of the nominative marking for all morphological types of stems:

(6) Standard Russian: NOM => Nanai Russian: NOM
Standard Russian: GEN/ACC => Nanai Russian: GEN/ACC~NOM as free variants

So we cannot consider the data of Nanai Russian as an evidence for chaotic erosion of Standard Russian animacy-driven split in bilinguals' speech. In particular, we do not attest genitive-like forms or dedicated accusative ones in the contexts in which nominative-like forms are expected in Standard Russian. Only one such example is attested (7).

(7) А у них *матери/* <=мать> давно хоронили/

The genitive-like form *матери* instead of the expected nominative-like *мать* can be interpreted here as the case of overgeneralization of semantic animacy-split to the non-appropriate morphological stem types. However this example is unique[6]. In outline, the Nanai System copies consistently not only the semantic distinction attested in Standard Russian, but also the formal split between different stem types.

2.3. The sample of DO-contexts

Taking into account the general picture presented in Table 2, in the remaining part of the paper I analyze only stems for which free variation between nominative forms and dedicated accusative or genitive ones is potentially expected in the Nanai Russian system (i.e. for which the accusative form in Standard Russian is not nominative-like), namely:

- a-stems singular (both animate and inanimate);
- 0-stems singular, animate;
- plural stems, animate.

My corpus of the speech of VSG (15 texts, 1601 clauses, 1 h. 15 min.) gives a sample of 94 examples. All of them are examples of the stems used in DO-position (in Nominative or in Accusative[7]). All other stems in DO-position (for which the Nominative-like form is the only option in DO-position both in Nanai Russian and in Standard Russian) were excluded from the sample[8].

---

[6] In fact there are also some less clear examples with 0-stems masculine, such as the following one: *Мы один раз нашли/ сундýка*. Here it is not evident if we deal with Genitive (with untypical stress position, *сундукá* is expected) or with a phonetic variant of Nominative (conditioned by a more general tendency to open final syllables in Nanai). The second option seems to be more probable, because the same forms are also attested as subjects (*стоит сундýка*).

[7] Further I refer both to dedicated accusative forms (for a-stems) and genitive-like forms (for 0-stems and plural stems) as "Accusative" (ACC).

[8] The following contexts were also excluded: a) numeral phrases in DO-position; b) negative contexts (because of the possible contamination with the Genitive-of-Negation construction).

3. NOM~ACC variation in DO-position: relevant factors

3.1. Animacy and definiteness: irrelevant

Semantic features of object which are expected to trigger the choice between different DO markers in languages of the world do not reveal any statistically significant correlations with NOM~ACC marking in our Nanai Russian data. In particular, the parameters that are relevant for DOM in Nanai and in Northern Russian dialects do not play any role in the Nanai Russian system.

3.1.1. Animacy, human vs. non-human distinction

Nanai Russian follows the same animacy distinction as Standard Russian (see 2.2). However there are no statistical correlations with animacy scale (cf. 1.1 above) within the pool of NOM~ACC free variation.

Table 3 presents the distribution of different types of objects on animacy scale for a-stems (for all other types the variation is possible only for animate objects, see above). The slight differences between NOMs and ACCs are not statistically significant.

Table 3. NOM~ACC variation and animacy: a-stems

	nom	acc
inanimates	78% (21)	73% (24)
animal/product (fish)[9]	11% (3)	9% (3)
animals	0% (0)	9% (3)
collectives	4% (1)	0% (0)
humans	7% (2)	0% (0)
proper (human) names	0% (0)	9% (3)
total	100% (27)	100% (33)

Table 4 contains the data on distribution between animals and humans for all inflection types. There is no significant correlation either.

Table 4. NOM~ACC variation and human/non-human distinction: all stems

	nom	acc
animals	31% (4)	27% (4)
humans	69% (9)	73% (11)
total	100% (13)	100% (15)

3.1.2. Definiteness, specificity

Definiteness and specificity of the object do not play a role either. Table 5 shows that the proportions of non-specific indefinites and of specific indefinites are slightly

[9] The uses of the word *рыба* 'fish' which is used in two senses ("animal" and "inanimate", fish-meal) and which is very frequent in our texts were counted separately.

larger for NOMs than for ACCs, however this difference is not statistically significant. Cf. example (8) with the definite object marked by NOM:

(8) Она же снимает/ *эта берёста*

Table 5. NOM~ACC variation and definiteness

	nom	acc	2-tailed exact Fisher test
definite	34% (14)	48% (24)	ns, p=0.2057
specific indefinite	17% (7)	12% (6)	
non-specific indefinite	49% (20)	40% (20)	ns, p=0.5246
total	100% (41)	100% (50)	

3.2. Predicate form: irrelevant

Unlike nominative objects in Russian dialects those of Nanai Russian have no predisposition towards infinitive clauses. There are only two such examples in our sample, cf. (9).

(9) Вот *такая* сделать/ на доски — три ряд\.

3.3. Word order and information structure

Table 6 shows the word order distribution in clauses with Nominative vs. Accusative objects. Cf. examples (10) and (11) with NOM:

(10) *Полный нарта* нагрузили/ тащили (OV)

(11) Берет опять *газета*/ (VO)

The percentage of OV-uses is a bit higher for NOMs than for ACCs (like in Russian Northern dialects), but the difference is not significant.

Table 6. NOM~ACC variation and word order

	nom	acc	2-tailed exact Fisher test
OV	68% (26)	50% (23)	p=0.1201, ns
VO	32% (12)	50% (23)	

The data on general distribution of topics / foci is not significant either, see Table 7.

Table 7. NOM~ACC variation and information structure: % of focused objects

	nom	acc	2-tailed exact Fisher test
topic	33% (13)	50% (23)	p=0.1270, ns
focus	67% (27)	50% (23)	

Still, a significant trend to Nominative marking is attested with the more subtle class of left-dislocated objects in focus position, as in (12). See Table 8.

(12) Этой кричит\ так: Оооо\! Это значит *медведь*$_{foc}$\ дед везет$_{foc}$

Table 8. NOM~ACC variation and information structure: % of objects in left-dislocated focus position

	nom	acc	2-tailed exact Fisher test
left-dislocated foci	39% (16)	18% (9)	p=0.0340
others	61% (25)	82% (41)	

Notably, indirect objects can be also marked by NOM in clauses with a non-standard information structure (though to a lesser extent), as mentioned above (2.1). This is an argument for possible interpretation of this DOM pattern as a part of more general syntactic strategy of information structure marking.

3.4. Formal features: inflection type

Factors which are the most relevant for the choice between NOM vs. ACC in DO-position are morphological and not semantic. This is a possible argument for the hypothesis of under-acquisition of the Standard Russian system.

There is not enough data to postulate a correlation of DOM with a declension type (a-stems singular vs. 0-stems singular vs. plural stems), as shown in Table 9.

Table 9. NOM~ACC variation and declension type

	nom	acc
a-stems sg	27	33
plural stems	5	7
0-stems masc. sg	7	0–2?

Still, a significant correlation with accentual types is attested within the most numerous a-stem class. The stems that have stress on the case affixes in ACC and in NOM (*ед-а*, *ед-у*) or at least in one of these forms (*голов-а*, *голов-у*) tend to take ACC in DO position. The stems with unstressed case affixes in ACC and NOM (*рыб-а*, *рыб-у*) tend to take NOM. See Table 10.

Table 10. NOM~ACC variation and accentual type: a-stems

	nom	acc	2-tailed exact Fisher test
case-affix unstressed	89% (24)	61% (20)	p=0.0189
case-affix stressed	11% (3)	39% (13)	

This rule can be reformulated in a following way. The stems with a higher degree of perceptive distinctiveness between NOM and ACC save the same opposition as in Standard Russian. The stems with a lower degree of perceptive distinctiveness between NOM and ACC lose the opposition between these forms in the under-acquired system of bilingual speakers, so the expansion of Nominative in DO-position is attested for the second type stems, rather than for the first type.

4. Discussion

Differential object marking attested in the Russian Speech of Nanai-Russian bilingual speakers presumably can have the following potential sources:

a) DOM pattern in Nanai;
b) DOM pattern in dialectal substrate of the local Russian variety;
c) incomplete acquisition of Standard Russian system among bilingual speakers.

Table 11 shows the results of detailed comparison of factors relevant for DOM in Nanai Russian with those relevant for DOM in Nanai and in Russian dialects.

Table 11. DOM in Nanai Russian, in Nanai and in Russian dialects

	Nanai Russian	Nanai (Oskolskaya, Stoynova 2017)	Russian Northern Dialects (Ron'ko 2017)
% of NOM's in DO-position	44% (in competing contexts)	52%	?
animacy	−	−	+ (inanimate)
definiteness	−	+ (indefiniteness)	+
word order	± (left—dislocated focus)	−	+ (OV)
information structure	+ (left—dislocated focus)	+ (−topic)	+ (focus)
predicate form	−	−	+ (infinitives)
inflection type	+ (accentual type)	0	?

The following conclusions can be made.

1) DOM pattern in Nanai Russian is not similar to the pattern attested in Russian dialects. The most sufficient structural difference is that in Nanai Russian DOM has no connection to infinitival constructions. At the same time there are no evident historical preconditions for such influence. So we estimate this potential source as very dubious.

2) DOM pattern in Nanai Russian reveals more similarity with the one attested in Nanai. However a) this similarity concerns the parameters which are not specific for Nanai, but rather typical of DOM in the languages of the world; b) DOM pattern in Nanai Russian has features which have no parallels in Nanai (cf. the correlation with morphological type of stem). So the morphosyntactic borrowing from Nanai can be estimated as one of sources of DOM in Nanai Russian, but not as the only one.

3) The hypothesis of incomplete acquisition seems to be probable. This is not the case of a chaotic set of occasional "errors" in L2 in the process of learning. The data we deal with present a rather clear consistent stable system. Moreover, this non-standard variety of Russian is near-native for VSG and nowadays it is her dominant language. So it is more accurate to describe the case as non-standard acquisition, rather than incomplete one. One can assume that the DOM pattern in Nanai Russian emerges as a systematic reinterpretation of the Standard Russian system in the specific situation of language contact and a lack of L2 input.
   a) Optional Nominative marking is added to the Standard Russian split Accusative marking system without breaking the initial system. This option itself can be interpreted as a result of a direct Nanai influence.
   b) Nominative penetrates more intensively into the parts of the noun paradigm that are more difficult to acquire. It covers most of all the stems with unstressed case markers for which the perceptive difference between Accusative and Nominative is minimal.
   c) Nominative marking of direct object can be brought into correlation with a more general trend to Nominative coding of non-standard information structure attested in Nanai Russian. This has no clear prototype in Nanai.

4) Some other non-standard uses of Nominative are attested in Nanai Russian beyond DO-position. These are mostly the contexts in which the nominative (=unmarked) form is used in Nanai. It is interesting if these uses in Nanai Russian are regulated by the stem accent type, in the same way as DO uses of Nominative (cf. 3) above). If yes, it would mean that the factor of interference (2) and the factor of non-standard acquisition or reinterpretation of the Russian system (3) work separately at different levels: 1) the Nanai system provides possible non-standard contexts for Nominative (interference) and then 2) the Russian system provides appropriate stems for Nominative (non-standard acquisition of Russian). Unfortunately, by the moment I have not enough data to test it.

References

1. *Aissen J.* (2003), Differential object marking: iconicity vs. economy, Natural Language & Linguistic Theory, 21(3), P. 435–483.
2. *Avrorin V. A.* (1948), Outline of Nanai Syntax. Direct Object [Očerki po sintaksisu nanajskogo jazyka. Prjamoe dopolnenie], Leningrad, Učpedgiz.
3. *Bossong G.* (1985), Differentielle Objektmarkierung in den neuiranischen Sprachen, Tübingen, Narr.
4. *Cox C.* (2011), Corpus Linguistics and Language Documentation: Challenges for Collaboration, Newman J., Baayen R. H. and Rice S. (Eds.), Corpus Linguistics: An International Handbook. Volume 1. Berlin — New York, Walter de Gruyter, P. 239–264.
5. *Dalrymple M., Nikolaeva I.* (2011), Objects and information structure, Cambridge, Cambridge University Press.
6. *Iemmolo G.* (2010), Topicality and differential object marking: Evidence from Romance and beyond, Studies in Language, 34(2), P. 239–272.
7. *Malchukov A.* (2008), Animacy and Asymmetries in Differential Case Marking, Lingua, 118(2), P. 203–221.
8. *Malchukov A., de Swart P* (2009), Differential case marking and actancy variation, Malchukov A., Spencer A. (Eds.), The Oxford Handbook of Case, Oxford, OUP, P. 290–303.
9. *Markova N. V.* (1989), Non-standard ways of expressing subject and object in Onega dialects and their history [Dialektnye sposoby vyraženija sub'jekta i ob'jekta v Onežskix govorax i ix istorija]. PhD Thesis, M.
10. *Mosel U.* (2014), Corpus Linguistic and Documentary Approaches in Writing a Grammar of a Previously Undescribed Language, Language Documentation & Conservation, 8, P. 135–157.
11. *Oskolskaya S. A., Stoynova N. M.* (2017), Differential object marking in Nanai [Differencirovannoje markirovanije ob'jekta v nanajskom jazyke], Acta Linguistica Petropolitana, XIII(3), P. 336–370.
12. *Ostler N.* (2008), Corpora of less studied languages, Lüdeling A. and Kytö M. (eds.), Corpus Linguistics: An International Handbook. Volume 1. Berlin — New York: Walter de Gruyter, P. 457–483.
13. *Ron'ko R.V.* (2017), Nado korova doit'! Nominative object in Northern Russian dialects [Nominativnyj ob'jekt v severnorusskix dialektax], Acta Linguistica Petropolitana, XIII(3). P. 244–264.
14. *Silverstein M.* (1976), Hierarchy of features and ergativity, Dixon R. M. W. (Ed.), Grammatical Categories in Australian Languages, Canberra, Australian Institute of Aboriginal Studies Publications, P. 112–171.
15. *Vinogradov I.* (2016) Linguistic corpora of understudied languages: do they make sense?, Káñina, 40(1).
16. *Witzlack-Makarevich A., Seržant I.* (2017), Differential argument marking: Patterns of variation, Seržant I., Witzlack-Makarevich A. (Eds.), The Diachronic Typology of Differential Argument Marking, Berlin, Language Science Press, P. 1–45.

ИНТЕРПРЕТАЦИЯ РУССКИХ МЕСТОИМЕНИЙ В КОНТЕКСТАХ КОНТРФАКТИЧЕСКОГО ТОЖДЕСТВА: ОПЫТ КОРПУСНОГО ИССЛЕДОВАНИЯ[1]

Тискин Д. Б. (daniel.tiskin@gmail.com)
Санкт-Петербургский государственный университет

В статье предпринимается попытка корпусного анализа семантики русских личных и притяжательных местоимений в интенсиональных контекстах (на примере контекстов контрфактического тождества). Задача исследования состояла в том, чтобы определить, способны ли местоимения различных типов интерпретироваться *de se* или *de re* в таких контекстах и какая из интерпретаций предпочтительна.

Контекстами контрфактического тождества называются синтаксические позиции, находящиеся в сфере действия модификатора или клаузы, вводящей ирреальное условие, касающееся тождества тех или иных нетождественных в действительности индивидов (ср. *на твоём месте*, англ. *if I were you*). В таких контекстах местоимение может обозначать реальную личность (как в *Я бы на их месте таких должников, как я, в хвост и гриву гоняла*; *de re*) или же ирреальную (*Я бы на их месте поставил парочку шалашиков в любом приглянувшемся мне месте*; *de se* — тот, с чьей точки зрения рассматривается ирреальная ситуация).

На материале ГИКРЯ (около 20 млрд словоупотреблений) мы показываем, что местоимения *я* и *мой* допускают как интерпретацию *de re,* так и интерпретацию *de se*, но первая предпочтительна; что возвратное местоимение *себя* также допускает обе интерпретации, но предпочтительнее *de se*; что возвратное притяжательное местоимение *свой* безысключительно интерпретируется *de se*. Кроме того, сделаны некоторые квалитативные наблюдения, касающиеся идентификации атомарного индивида с множественным, как в *я бы на вашем месте не стала морочить себе голову, вы молодые люди у вас ещё всё впереди*.

Ключевые слова: анафора, интернет-корпус, контрфактическое тождество, русский язык, *de re*, *de se*

[1] Исследование выполнено при поддержке РФФИ, проект № 18-011-00895.

THE INTERPRETATION OF RUSSIAN PRONOUNS IN COUNTERIDENTITY CONTEXTS: A CORPUS STUDY

Tiskin D. B. (daniel.tiskin@gmail.com)
Saint Petersburg State University

This paper is a first step towards a corpus-based description of the semantics of Russian pronouns in intensional contexts. Having justified the use of corpus in (formal) semantic research, I delineate a particular issue within the topic: whether a given pronoun is interpreted *de se* or *de re* in counteridentity contexts.

A counteridentity context is a clause within the scope of a counterfactual (clause or adverbial) that affects the identity of a real individual, e.g. *if I were you*, *were I you*, etc. If a pronoun such as *I*, *my* or the Russian reflexive possessive *svoj* is used in such a context, two options are theoretically possible: either it picks out the speaker's real self (*de re*), or it refers to the identity assumed by the speaker in the contrary-to-fact situations introduced by the counterfactual (*de se*).

Using data from the GICR corpus (approx. 20 billion tokens), I show that for the Russian first-person singular pronoun *ja* and its corresponding possessive *moj*, *de se* reference is possible but *de re* interpretation is more frequent. The opposite holds for the reflexive *sebja*, whereas *svoj* is interpreted *de se* with no exception. Special attention is paid to situations where more than one referential strategy is possible. The paper concludes with a couple of observations relevant for the future formal accounts of *de se* reference.

Keywords: anaphora, counteridentity, *de re*, *de se*, Russian, Web corpus

1. Интерпретация местоимений в косвенных контекстах

Со времён Г. Фреге [Frege 1892] сентенциальные комплементы некоторых предикатов, таких как модальные предикаты или глаголы пропозициональных установок, рассматриваются как источники «референциальной непрозрачности»: значение предложения оказывается зависимым не только от значений (денотатов) входящих в него выражений и способа их сочетания, но и от того, что Фреге назвал «смыслом», а Р. Карнап [Carnap 1947] — интенсионалом. Референциальная непрозрачность, однако, в значительной части случаев является лишь одной из возможностей наряду с «прозрачностью» — такой интерпретацией выражения, находящегося в сфере действия интенсионального оператора, как если бы никакого оператора не было. Так, (1) может означать 'Петрович думает: все лингвисты, кто бы они ни были, бездельники' или же 'Есть группа лиц, которых Петрович (знает и) считает бездельниками, а ещё эти лица как раз и составляют множество всех лингвистов'. В последнем случае интерпретация оказывается референциально прозрачной, или *de re*.

(1) *Петрович думает, что все лингвисты — бездельники.*

Специфическая разновидность такой неоднозначности возникает, когда в сферу действия оператора попадает выражение, способное интерпретироваться *de se*. Под интерпретацией *de se* [Castañeda 1966; Lewis 1979] понимается такая, при которой носителю пропозициональной установки приписывается мысль «от первого лица». Так,

(2) *Наш кандидат$_1$ думает, что он$_1$ победит.*

может означать, что кандидат имеет мысль, выразимую как «**Мне** обеспечена победа». С другой стороны, предложения типа (2) допускают и интерпретацию *de re*; она была бы истинна, в частности, тогда, когда наш кандидат, изучая отчёт о кампании неизвестного ему лица, думал бы: «**Этот человек** победит», — тогда как на самом деле отчёт касался бы его собственной кампании. Несмотря на то что условия истинности интерпретации *de se* в случаях типа (2) сильнее, чем у *de re* (истинность *de se* гарантирует одновременную истинность *de re*, но не наоборот), Percus and Sauerland [2003a] убедительно показывают, что *de se* представляет собой отдельное прочтение подобных предложений.

Имеются выражения, для которых интерпретация *de se* обязательна. К ним относятся pro ([Chierchia 1989]; но см. возражения в [Cappelen and Dever 2013]) и, возможно, (некоторые) логофорические местоимения [Schlenker 2003; но см. данные в Pearson 2015], а также предикаты-эгоцентрики типа *вкусно* [см. Pearson 2013; Падучева 2017] и эвиденциальные показатели [Korotkova 2017].

2. Контексты контрфактического тождества

Дж. Лакофф [Lakoff 1970] обратил внимание на примеры, где в сфере действия предиката установки оказывается местоимение, совпадающее с субъектом предиката установки:

(3) *I dreamed that I was playing the piano.*
'Мне снилось, что я играю на пианино'

(4) *I dreamed I was Brigitte Bardot and I kissed me.*
'Мне снилось, что я Брижит Бардо и что я поцеловал меня'

Лакофф указывает, что в (3) второе *I* может означать как личность, с чьей точки зрения во сне воспринимается мир (тогда говорящий видит пианино с точки зрения сидящего за ним пианиста и т. п.; *de se*), так и личность говорящего при взгляде на неё со стороны (говорящий видит себя самого за пианино со стороны; *de re*). Пример (4) может означать, что говорящий «в теле» Брижит Бардо, с чьей точки зрения он видел сон, поцеловал говорящего, которого видел со стороны (и только в этом случае (4) не нарушает требования теории связывания, согласно которому следовало бы употребить *myself* вместо *me*). Обратное невозможно: у (4) нет интерпретации, при которой во сне некто в теле говорящего целует говорящего, заключённого в тело Брижит Бардо (см. об этом [Percus and Sauerland 2003b]).

К примерам Лакоффа примыкают примеры, где ирреальное тождество вводится оборотами типа *if I were you* или *на твоём месте*. А. Арреги [Arregui 2007] отмечает, что в контексте *if I were you* допускается референция *me* к реальной, а не контрфактической личности говорящего, тогда как *if you were him/her* и *if (s)he were you* практически не допускают такой возможности для, соответственно, *you* и *him/her*. Наша выборка для русского языка (см. §4) сформирована с учётом этого обстоятельства; впрочем, в ГИКРЯ встречаются единичные примеры наподобие (5)–(6).

(5) *…но он сказал, что если бы он был мной, то поступил, как и я — развелся бы с ним.*

(6) *«На моём месте ты бы не стал тусить с таким, как ты?»*

Наконец, К. Кауф [Kauf 2016, 2017] обращает внимание на то, что отмеченные Арреги ограничения — общие у контрфактического тождества с примерами типа (4). Кауф предполагает, что языки, в которых, как в русском, типичный зачин для советов ‘на твоём месте’ выглядит не так, как клауза со значением ‘если бы я был тождествен тебе’, свидетельствуют о неоднозначности *if I were you* в английском; в значении совета *if I were you* не означает отождествления говорящего с адресатом, пусть даже и контрфактического.

3. Мотивация исследования

Хотя в последние десятилетия исследования референции местоимений в косвенных контекстах приобрели популярность, корпусные методы, насколько нам известно, в таких исследованиях до сих пор использовались незначительно. Действительно, использование корпусов сталкивается здесь с двумя проблемами:

(a) корпусные данные не сопровождены семантическим комментарием, и в их интерпретации исследователь вынужден полагаться на суждение лиц, не являющихся авторами высказывания;
(b) исследуемые явления, такие как чтения *de re* для местоимений, способных иметь интерпретацию *de se*, достаточно редки в естественной речи, так что в корпусе может не оказаться достаточного числа примеров.

Вторая проблема оказалась отчасти решена с появлением сверхбольших корпусов, таких как, в частности, Генеральный интернет-корпус русского языка (ГИКРЯ, http://webcorpora.ru). Корпус, насчитывающий около 20 млрд словоупотреблений, уже достаточно велик и представителен в интересующем нас отношении, чтобы даже несложные запросы к нему могли быть информативны (т. е. содержать релевантные данные и притом в количестве, достаточном хотя бы для элементарных статистических выкладок).

Частичное решение первой проблемы состоит в том, что работающий с корпусом исследователь сам является носителем исследуемого языка, а в случае необходимости может обратиться к расширенному контексту, в т. ч. и путём перехода на интернет-страницу, откуда взят пример (там, где это возможно).

Разумеется, отсутствие тех или иных примеров в корпусе не означает их неприемлемости, но сравнение частот и само по себе оказывается информативным. Наконец, выбор русского языка, помимо сказанного в конце §2, представляет интерес потому, что русский отличается от английского языка, до сих пор подробнее всего описанного в интересующем нас отношении, наличием возвратного притяжательного местоимения.

4. Материал исследования

Материалом исследования послужили извлечённые из ГИКРЯ контексты, в которых личное, возвратное или притяжательное местоимение выступает в сфере действия контрфактического условия, семантика которого затрагивает тождество индивидов, называемых данными местоимениями. Принимаемая идентичность могла варьировать, но множество «реципиентов» новой идентичности мы ограничили говорящим — референтом местоимения *я*. Поскольку нас интересовало поведение местоимений в контекстах, затрагивающих именно тождество их референтов, то, к примеру, в контексте условия *на твоём месте я бы…* рассматривалось поведение местоимений *я*, *мой*, *ты*, *твой*, *мы*[2], *наш*, *вы*[3], *ваш*, *себя*, *свой*, но не *он*, *она*, *оно*, *они*; в контексте условия *на её месте я бы…* вместо *ты*, *твой*, *вы*, *ваш* рассматривались *она*, *её*; аналогично для других типов условий. Не учитывалась нулевая анафора, как в (7), где опущено подлежащее, восстановимое по согласованию формы *докажете*.

(7) *Но на практике, я бы на Вашем*$_1$ *месте рисковать не стала, слепо думая, что* $\emptyset_1$ *докажете свою невиновность*[4]

Кроме того, не учитывался субъект главной клаузы — *я* (однако в случае вложения клауз субъекты всех придаточных, находящихся в сфере действия контрфактического условия, учитывались).

Для каждого рассматриваемого местоимения мы определяли, как устанавливается референция местоимения: в мире произнесения (т. е. в действительном мире, где *я* — это говорящий, а *ты*, *она* и пр. — отличные от говорящего индивиды), как в (8), или же в соответствующих ирреальному условию мирах, где личность говорящего «перенесена» в личность другого индивида, как в (9).

(8) *я бы на твоём месте сделал <u>себе</u> такой юзерпик для комментов)*

---

[2] В случаях, когда в условии выступало местоимение единственного числа, мы учитывали и местоимения множественного числа, если множество, являющееся их референтом, включало референт местоимения ед. ч. Случаи, когда местоимение мн. ч. отсылало к совокупности, включающей как говорящего, так и референта местоимения ед. ч., однако, исключались; ср.: *Так что я бы на вашем месте составил для нашего города не одну заявку, а 3–4 сразу* (видимо, говорящий и адресат живут в одном городе).

[3] Местоимения *вы* и *Вы* нами не различались.

[4] Этот и все последующие нумерованные примеры взяты из ГИКРЯ. Примеры из ГИКРЯ могут быть фрагментарными. Мы восстанавливаем пример целиком там, где доступен исходный текст страницы.

(9) *Я бы на твоём месте уже бы послала бы себя куда подальше.*

В (8) *себе* относится к личности говорящего в альтернативном мире, где он «встаёт на место» адресата, тогда как в (9) *себя* отсылает к самой говорящей: её похвала адресату состоит в том, что именно она, а не адресат, непроста в общении, но адресату удаётся с нею уживаться.

Подсчёты, которые были сделаны, могут быть использованы двояко:

(a) для выяснения того, с какой частотой (и в каких случаях чаще) то или иное местоимение используется с актуальной референцией и с какой — со «смещённой» референцией, соответствующей контрфактическим альтернативам;

(b) для выяснения того, какое из потенциально синонимичных местоимений, конкурирующих в данной позиции, чаще выбирается говорящими.

Случаи, которые мы имеем в виду в (b), — это конкуренция *я* и *себя*, а также *мой* и *свой* за роль единицы, обозначающей действительную личность говорящего (sp-r) и за роль единицы, обозначающей личность говорящего в альтернативных мирах (sp-c).

5. Результаты

Наша основная выборка 1 состоит из контекстов, в которых встречается последовательность *на* X *месте я* или *я бы на* X *месте*, где X ∈ { *вашем*, *его*, *её*, *их*, *твоём* }[5].

5.1. Семантика местоимений

Следуя описанной выше процедуре включения контекстов в рассмотрение, мы получили выборку следующего состава.

В **Таблице 1** подчёркнуты числа, соответствующие вхождениям тех же местоимений, что и местоимение в контрфактическом условии.

Мы объединяем личные и связанные с ними притяжательные местоимения, условно именуя их вместе «(от)личными». Из **Таблицы 2** видно различие между возвратным местоимением *себя* и соответствующим ему притяжательным *свой*, однако допустимо и объединить их (ср. термин *reflexive possessive*); эту группу мы называем «(от)возвратными». С учётом этой конвенции данные **Таблицы 2** могут быть обобщены следующим образом.

Столь высокая доля реальной референции для личных местоимений отчасти обусловлена тем, что мы учитывали и местоимения 2 и 3 лица, отсылающие к упоминаемым в контрфактическом условии индивидам. Тем не менее, для *я* доля реальной референции также высока (82,8 %).

[5] В данной работе мы отвлекаемся от того обстоятельства, что *себя* в значительной части случаев выступает в составе устойчивых выражений *вести себя*, *чувствовать себя*, *покончить с собой*.

Таким образом, для личных местоимений не исключена интерпретация *de se*, связанная с «новыми» идентичностями, которые приносит с собой косвенный контекст, но предпочитается всё же интерпретация *de re*, связанная с первичным дейксисом. Возвратное местоимение в примерно равной степени допускает обе стратегии[6], а возвратное притяжательное **всегда** интерпретируется *de se* — ассоциируется с ближайшим доступным дейктическим центром (в данном случае — вводимым ирреальным условием), что приводит к монополии вторичного дейксиса там, где он вообще возможен.

Таблица 1. Референция исследованных местоимений в выборке 1 (подсчёт контекстов)

В условии	Референция	Местоимения														Σ
		ваш	*вы*	*его*	*мой*	*мы*	*наш*	*он*	*она*	*они*	*свой*	*себя*	*твой*	*ты*	*я*	
ваш	реальная	33	50	0	18	7	1	0	0	0	0	7	0	0	47	**163**
	смещённая	0	0	0	3	0	0	0	0	0	36	42	0	0	9	**90**
его	реальная	0	0	2	4	11	4	11	0	0	0	29	0	0	71	**132**
	смещённая	0	0	0	8	0	0	0	1	0	34	57	0	0	17	**117**
её	реальная	0	0	0	0	2	1	0	2	0	0	2	0	0	12	**19**
	смещённая	0	0	0	3	0	0	0	0	0	3	7	0	0	2	**15**
их	реальная	0	0	0	5	14	5	0	0	4	0	20	0	0	33	**81**
	смещённая	0	0	0	4	0	0	0	0	0	26	39	0	0	6	**75**
твой	реальная	2	0	0	5	1	0	0	0	0	0	7	8	17	20	**60**
	смещённая	0	0	0	1	0	0	0	0	0	22	17	0	0	4	**44**
Σ	реальная	**35**	**50**	**2**	**32**	**35**	**11**	**11**	**2**	**4**	**0**	**65**	**8**	**17**	**183**	**455**
	смещённая	**0**	**0**	**0**	**19**	**0**	**0**	**0**	**1**	**0**	**121**	**162**	**0**	**0**	**38**	**341**

Таблица 2. Референция различных групп местоимений в зависимости от семантики основы в выборке 1. Для всех пар столбцов, кроме «личные» vs. «притяжательные (от личных)», $p \ll 0.001$[7]

референция	личные	притяжательные (от личных)	возвратное (*себя*)	притяжательное (от возвр.: *свой*)
реальная	302 (88,6%)	88 (82,2%)	65 (28,6%)	0 (0%)
смещённая	39 (11,4%)	19 (17,8%)	162 (71,4%)	121 (100%)

[6] Из таблицы 1 видно, что *себя* значимо чаще имеет реальную интерпретацию при *его* или *их* в условии, чем при *ваш*. Возможно, это связано с тем, что вторичный дейксис, связанный с точкой зрения собеседника, более естествен, чем связанный с точкой зрения третьего лица.

[7] Здесь и далее используется двусторонний вариант критерия χ^2.

Таблица 3. Референция личных/возвратных и притяжательных местоимений в зависимости от семантики основы в ГИКРЯ (выборка 1). $\chi^2 \approx 371{,}16$, $p \ll 0.001$

референция	(от)личные	(от)возвратные
реальная	390 (87,1%)	65 (18,7%)
смещённая	58 (12,9%)	283 (81,3%)

5.2. Выбор стратегии выражения

Обратимся теперь к проблеме выбора средств выражения в случаях, когда теоретически существует больше одной единицы, способной выразить требуемое значение в данной позиции. Это касается значений ‘sp-r’, ‘sp-c’ и соответствующих им посессивных значений.

Необходимо помнить, что, помимо собственно семантических ограничений на референцию местоимений, могут существовать дистрибутивные ограничения на их употребление. Так, если ограничить подсчёты придаточными предложениями (в смысле традиционной грамматики), доля *себя* и *свой* значительно падает везде, где она не была уже равна 0 (см. таблицу 4). Это, разумеется, связано с тем, что анафор[8] в нормальном случае должен быть связан в своём домене [Chomsky 1981; Тестелец 2001: 598], а потому *себя* и *свой* с интересующей нас референцией встречаются в придаточном либо если подлежащее этого придаточного — тоже местоимение из нашего набора (10) (в т. ч. опущенное (11)), либо если это подлежащее нулевое, как в (12)–(13) [см. Падучева 1985: 191–192].

(10) *…хотя я бы на их месте делала бы ставки, когда я споткнусь и заплетусь на своих 10-сантиметровых каблуках, которые одела в честь весны, чтобы показать свои красивые ноги.*

(11) *Я бы на её месте обиделся, если бы* ∅ *мог найти свои носки под диваном*

(12) *я бы на вашем месте стала бы думать, как* pro *свою экскурсию провести так, чтобы деткам*

(13) *Я бы на Вашем месте не педалировал особенно эту тему, чтобы* pro *не создавать о себе странное впечатление.*

[8] Заметим, впрочем, что в некоторых из наших примеров *свой* субстантивировано: *На его месте я бы плюнул на эти стандарты и снимал бы своё.*

Таблица 4. Выбор стратегии выражения sp-r, sp-c и соответствующих посессивных значений в ГИКРЯ (выборка 1)

Вся выборка							
sp-r		**sp-c**		**sp-r.poss**		**sp-c.poss**	
я	183 (74%)	*я*	38 (19%)	*мой*	32 (100%)	*мой*	19 (14%)
себя	65 (26%)	*себя*	162 (81%)	*свой*	0 (0%)	*свой*	121 (86%)
$p \ll 0.001$				$p \ll 0.001$			

Только придаточные							
sp-r		**sp-c**		**sp-r.poss**		**sp-c.poss**	
я	26	*я*	30	*мой*	2	*мой*	15
себя	0	*себя*	5	*свой*	0	*свой*	5
—				—			

Итак, хотя *я* и *себя* оба допускают интерпретацию ‘sp-r’ или же ‘sp-c’, они неодинаковы с точки зрения предпочтений: для ‘sp-r’ значимо чаще предпочитается *я*, для ‘sp-c’ — *себя*. В случае притяжательных местоимений неспособность *свой* выражать ‘sp-r’ приводит к безальтернативности *мой* в этом значении. Тот факт, что наличие или отсутствие в языке возвратных притяжательных местоимений приводит к сужению или, соответственно, расширению сферы употребления невозвратных, упоминается в литературе для французского [Charnavel 2009: 66] и русского [Déchaine and Wiltschko 2010] языков; здесь же изменение сферы употребления *мой* происходит в пределах одного языка в зависимости от пригодности той или иной позиции для употребления рефлексивов. Сравнительно большая доля *мой* в случае ‘sp-c’ в основном связана с его употреблением в придаточных, где *свой* невозможно употребить из-за большого синтаксического расстояния от антецедента (14), хотя есть и случаи выбора *мой* в ситуациях конкуренции (15).

(14) *Не знаю, согласятся ли со мной украинцы, но я бы на их месте гордился, если бы мой* [**свой*] *земляк был создателем первой ракетной техники*

(15) *Я бы на твоём месте думала бы как это сообщить моему* [ОК*своему*] *босу…))))*

5.3. Различия в контрфактических условиях?

В §2 мы упомянули рассуждения К. Кауф о различиях между *на твоём месте* и *если бы я был тобой*, ср. (16)–(17). Чтобы проверить, распространяются ли эти различия на поведение местоимений, мы сформировали отдельную выборку.

(16) *Если бы я была тобой, то в данной ситуации просто не знала бы, куда себя деть от ярости)*

(17) *будь я тобой, я бы на меня не надеялся.*

Выборка 2 состоит из контекстов с последовательностью *будь я* Y или *если бы я был*(*а*) Y, где Y ∈ { *вами*, *ей*, *ею*, *им*, *ими*, *тобой* }. Они существенно менее распространены, чем контексты, составившие выборку 1. Как и для выборки 1, наблюдается значимое различие между (от)личными и (от)возвратными местоимениями:

Таблица 5. Референция личных/возвратных и притяжательных местоимений в зависимости от семантики основы в ГИКРЯ (выборка 2). $\chi^2 \approx 4{,}23$, $p < 0.05$

референция	(от)личные	(от)возвратные
реальная	22	7
смещённая	8	11

Дальнейшие различия ввиду малого числа примеров оказались незначимы, хотя личные местоимения и *себя* и здесь демонстрируют обе стратегии выбора референта, а *свой* безысключительно получает смещённую референцию везде, где два типа референции различаются[9].

6. Заключительные замечания

Сверхбольшие корпусы делают возможными корпусные исследования проблем, ранее изучавшихся только с опорой на компетенцию носителя, в частности — при условии, что интерпретатором корпусных примеров выступает носитель, — ряда проблем формальной семантики. Данная статья представляет собой первый опыт корпусного исследования референции местоимений в контекстах контрфактического тождества в русском языке.

Как было показано выше, (от)личные и возвратное местоимения способны отсылать как к действительной, так и к контрфактической личности. Тем не менее, в случае говорящего *себя* имеет смещённую референцию значительно чаще, чем актуальную, и выбирается для выражения смещённой референции значительно чаще, чем для выражения актуальной. Возвратное притяжательное местоимение безысключительно отсылает к контрфактической личности, если находится в сфере действия оператора, вводящего таковую в рассмотрение.

Ряд вопросов остался вне нашего обсуждения. Так, мы не рассматривали случаев, когда глагол, согласующийся с *я* по ед.ч., имеет семантику,

[9] К числу случаев, где различия нет (и которые мы по возможности отбрасывали), принадлежат такие, в которых им или ею имеет нереферентный референциальный статус [Падучева 1985: 147–150; Тискин 2015]. Иногда для решения этого вопроса требуется обращение к более широкому контексту:

(i) *Хорошо, что я не летучая мышь*..
..ведь если бы я была ею, я не смогла бы ходить со своим любимым плеером.

несовместимую с единичностью субъекта. Такой случай в нашей выборке был, впрочем, только один[10]:

(18) *так что на вашем месте я б объединялся в Совет Ветеранов, хотящих выжить в условиях дикой конкуренции.*

Наконец, отметим, что самый факт приемлемости предложений с контрфактическими условиями *на вашем месте я…*, *на их месте я…*, *будь я вами* и пр., по-видимому, заставляет внести модификации в семантику для *de se*, основанную на идее У. В. О. Куайна и Д. Льюиса [Quine 1968; Lewis 1979] о **центрированных** мирах: обычно считается, что значением (придаточного) предложения в такой семантике является множество пар ⟨мир, индивид[11]⟩, однако возможность «переноса идентичности» говорящего на множество (референт *вы* или *они*) показывает, что в качестве вторых элементов пар могут выступать не только атомарные, но и множественные индивиды — мереологические суммы атомарных (в смысле [Link 1983]). Практически полный запрет на дистрибутивные клаузы типа *?*если бы я был каждым из вас* не является достаточным аргументом (он может быть синтаксическим), но в (18) множественности требует семантика.

Литература

1. *Arregui A.* (2007). *Being me, being you*: Pronoun Puzzles in Modal Contexts. In: E. Puig-Waldmüller (ed.), Proceedings of Sinn und Bedeutung 11. Barcelona: Universitat Pompeu Fabra. Pp. 31–45.
2. *Cappelen H., Dever J.* (2013). The Inesssential Indexical: On the Philosophical Insignificance of Perspective and the First Person. OUP.
3. *Carnap R.* (1947). Meaning and Necessity. University of Chicago Press.
4. *Castañeda H.-N.* (1966). 'He': A study in the logic of self-consciousness. Ratio, vol. 8, pp. 130–157.
5. *Charnavel I.* (2009). Linking Binding and Focus: on intensifying *son propre* in French. MA thesis, UCLA.
6. *Chierchia G.* (1989). Anaphora and attitudes *de se*. In: R. Bartsch, J. van Benthem and P. van Emde Boas (eds.), Semantics and Contextual Expression. Kluwer/Reidel. Pp. 1–31.
7. *Chomsky N.* (1981). Lectures Government and Binding: The Pisa Lectures. De Gruyter.
8. *Déchaine R.-M., Wiltschko M.* (2010). When and why can 1st and 2nd person pronouns be bound variables? Ms., University of British Columbia.
9. *Frege G.* (1892). Über Sinn und Bedeutung. Zeitschrift für Philosophie und philosophische Kritik, Bd. 100, S. 25–50.

[10] Ср., впрочем, и пример с множественностью **в составе** группы сказуемого: *на вашем месте я бы серьезно задумалась над состоянием и наполнением своих голов*.

[11] Т. е. «центр» мира, тот индивид в мире — первом компоненте пары, с которым отождествляет себя носитель пропозициональной установки.

10. *Kauf C.* (2016). Counterfactuals and (counter-)identity: The identity crisis of "If I were you". MA thesis, University of Göttingen.
11. *Kauf C.* (2017). An Analysis of Counteridenticals in Terms of Dream Reports. Presented at Sinn und Bedeutung 22.
12. *Korotkova N.* (2017). A novel route to 'de se'. Presented at the 18th Szklarska Poręba Workshop on the Roots of Pragmasemantics.
13. *Lakoff G.* (1970). Counterparts, or the Problem of Reference in Transformational Grammar. URL: https://files.eric.ed.gov/fulltext/ED022152.pdf.
14. *Lewis D.* (1979). Attitudes *de dicto* and *de se*. The Philosophical Review, vol. LXXXVIII, no. 4. Pp. 513–543.
15. *Link G.* (1983). The Logical Analysis of Plurals and Mass Terms: A Lattice-Theoretical Approach. In: R. Bäuerle, C. Schwarze and A. von Stechow (eds.), Meaning, Use and Interpretation of Language. Berlin/New York: de Gruyter. Pp. 303–323.
16. *Paducheva E. V.* (1985). The Utterance and Its Relation to Reality [Vyskazyvanie i ego sootnesyonnost' s deystvitel'nostyu]. Moscow: Nauka. [Падучева Е. В. (1985). Высказывание и его соотнесённость с действительностью. М.: Наука.]
17. *Paducheva E. V.* (2017) Egocentric items in language [Ègotsentricheskie yazykovye edinitsy]. Russian corpus grammar (http://rusgram.ru). [Падучева Е. В. (2017). Эгоцентрические языковые единицы. Материалы для проекта корпусного описания русской грамматики (http://rusgram.ru). На правах рукописи.]
18. *Pearson H.* (2013). A judge-free semantics for predicates of personal taste. Journal of Semantics, vol. 30, no. 1. Pp. 103–154.
19. *Pearson H.* (2015). The interpretation of the logophoric pronoun in Ewe. Natural Language Semantics, vol. 23, no. 2. Pp. 77–118.
20. *Percus O., Sauerland U.* (2003a). On the LFs of attitude reports. In: Proceedings of Sinn und Bedeutung 7. Pp. 228–242.
21. *Percus O., Sauerland U.* (2003b). Pronoun movement in dream reports. In: Proceedings of NELS, vol. 33. Pp. 265–284.
22. *Quine W. V. O.* (1968). Propositional objects. Crítica: Revista Hispanoamericana de Filosofía, vol. 2, no. 5. Pp. 3–29.
23. *Schlenker P.* (2003). A plea for monsters. Linguistics and Philosophy, vol. 26, no. 1. Pp. 29–120.
24. *Testelets Ya. G.* (2001). General Syntax: An Introduction [Vvedenie v obshchiy sintaksis]. Moscow: RSUH. [Тестелец Я. Г. (2001). Введение в общий синтаксис. М.: РГГУ.]
25. *Tiskin D. B.* Anaphora to predicates and the properties of Russian NPs [Anafora k predikatu i kharakteristiki russkoy imennoy gruppy]. In: E. Lyutikova, A. Zimmerling, M. Konoshenko (eds.), Typology of Morphosyntactic Parameters 2015. Moscow: MSPU. Pp. 334–352. [Тискин Д. Б. (2015). Анафора к предикату и характеристики русской именной группы. Е. А. Лютикова, А. В. Циммерлинг, М. Б. Коношенко (ред.) Материалы международной конференции «Типология морфосинтаксических параметров — 2015». М.: МПГУ. С. 334–352.]

THE CUES FOR RHETORICAL RELATIONS IN RUSSIAN: "CAUSE—EFFECT" RELATION IN RUSSIAN RHETORICAL STRUCTURE TREEBANK[1]

Toldova S. (toldova@yandex.ru)
NRU Higher School of Economics, Moscow, Russia

Pisarevskaya D. (dinabpr@gmail.com),
Kobozeva M. (kobozeva@isa.ru)
Institute for Systems Analysis FRC CSC RAS, Moscow, Russia

Vasilyeva M. (linellea@yandex.ru)
Lomonosov Moscow State University, Moscow, Russia

The purpose of the paper is to investigate cues signalling the relations between discourse units in Russian. Building a lexicon of discourse connectives is an indispensable subtask in many discourse parsing applications as well as an essential issue in theoretical researches of text coherence. In order to develop such a resource for Russian, we have conducted a corpus-based study of discourse connectives that were manually extracted from the Russian Rhetorical Structure Treebank (Ru-RSTreebank). The Treebank includes 79 texts annotated within the RST framework [Mann, Thompson 1988]. In order to provide a deeper analysis of connectives in Russian, we focus on causal relations only, namely, the 'Cause-Effect' relation. Some of the connectives (primary connectives) are enumerated in grammars and dictionaries. They primarily mark the intra-sentential relations. However, there is an expansive class of less grammaticalized items (secondary connectives) that have received less attention till now. Some of them are based on content words (e.g. по причине 'for the cause'). Secondary connectives often serve as linking devices for inter-sentential relations.

We suggest a scheme for connectives annotation for Russian. We specify the basic patterns that can be used for less-grammaticalized connectives mining in an unannotated corpus. Besides, we provide the comparison of two classes of connectives (primary vs. secondary ones). Our research has shown that these two classes differ in their properties. There is a statistically significant difference between them with respect to the nucleus/satellite position, intra- vs. inter-sentential relations and some others.

Keywords: discourse analysis, rhetorical structure theory, discourse connectives, corpus linguistics, corpus annotation

[1] The study was funded by RFBR according to the research project № 17-29-07033

МАРКЕРЫ РИТОРИЧЕСКОГО ОТНОШЕНИЯ «ПРИЧИНА — СЛЕДСТВИЕ» В РУССКОМ ЯЗЫКЕ НА МАТЕРИАЛЕ КОРПУСА РИТОРИЧЕСКИХ СТРУКТУР

Толдова С. (toldova@yandex.ru)
НИУ ВШЭ, Москва, Россия

Писаревская Д. (dinabpr@gmail.com),
Кобозева М. (kobozeva@isa.ru)
Институт системного анализа ФИЦ ИУ РАН, Москва, Россия

Васильева М. (linellea@yandex.ru)
МГУ имени М. В. Ломоносова, Москва, Россия

Создание лексикона дискурсивных коннекторов является одной из актуальных задач при разработке систем автоматического анализа дискурса. Описание коннекторов также играет немаловажную роль в теоретических исследованиях связности текста. В целях создания соответствующего лексикона для русского языка мы провели корпусное исследование коннекторов, выделенных экспертами в корпусе Ru-RSTReebank. Этот корпус представляет собой 79 научно-популярных и новостных текстов, размеченных в терминах теории риторических структур [Mann, Thompson 1988]. Вопрос о том, как устроен класс маркеров риторических отношений в русском рассматривается на примере каузальных отношений, в частности, на примере отношения "причина-эффект". Некоторые коннекторы (первичные коннекторы) представлены в грамматиках и словарях. Как правило, они маркируют связи внутри предложения. Однако существует достаточно обширный класс менее грамматикализованных коннекторов (вторичные коннекторы), которые исследованы в меньшей степени. В частности, в качестве коннекторов используются конструкции с полнозначными лексическими единицами (например, по причине). Многие из таких коннекторов маркируют связи между предложениями и дискурсивными единицами большего объема. Таким образом, настоящая работа посвящена анализу коннекторов, которые обеспечивают связь между дискурсивными единицами в русском языке. Особое внимание уделяется менее грамматикализованным коннекторам, в том числе коннекторам, обеспечивающим связность на меж-сентенциальном уровне. В работе мы предлагаем схему описания маркеров риторических отношений, разработанную на основе проведенного анализа, описываем основные модели образования свободных конструкций, с помощью которых список коннекторов может быть расширен с использованием неразмеченного корпуса текстов. Также в статье рассматриваются результаты сравнения двух классов коннекторов (первичных и вторичных). Между данными классами наблюдается статистически значимая разница в отношении ряда признаков, таких, например, как положение

внутри ядра/сателлита, тенденция к маркированию внутрисентенциальных vs. меж-сентенциальных отношений и др.

Ключевые слова: дискурсивный анализ, теория риторических структур, дискурсивные коннекторы, корпусная лингвистика, корпусная разметка

1. Introduction

The analysis of discourse structure is a challenging issue for linguistic theory. It plays an important role in many high-level NLP applications, such as text summarization [Louis et al. 2010], sentiment analysis [Voll and Taboada 2008], question answering [Ferrucci et al. 2010], argumentative discourse analysis [Galitskij et al. 2018] and others. Discourse parsing presupposes establishing the relations between coherent text spans. In many approaches, the identification of these relations relies on detecting special lexical clues. Thus, constructing a lexicon of discourse connectives is an essential task.

In this paper, we present a corpus study of discourse connectives for Russian from the perspective of constructing such a lexicon. As a source of data, we use the pilot Russian RST Treebank, built in 2017[2]. To provide an in-depth analysis of connectives, we focus on causal relations with a special emphasis on ‘Cause-Effect’ relation.

We consider two basic classes of connectives, namely, primary vs. secondary ones depending on whether they are registered in Russian grammars or not (cf. primary vs. secondary connectives distinction in [Rysová M., Rysová K. 2014]).

The first class includes among others different functional words such as conjunctions, prepositions etc. They have been studied for many years [Shvedova ed. 1980]; [Pekelis 2014, etc.]. However, the role of these connectives as signals of rhetorical relations within the RST framework still remains under-investigated. Besides, little attention was paid in the literature to less grammaticalized items that can serve as clues for inter-sentential relations.

As the result of our study, we provide the analysis of connectives with a special emphasis to less-grammaticalized items. We suggest a scheme for connectives annotation for Russian based on our corpus study with due consideration of other approaches [Roze et al. 2012]; [Stede, Umbach 1998]; [Mírovský et al. 2017] etc. We specify the basic patterns for less-grammaticalized connectives. These patterns can be used for extracting new connectives from an unannotated corpus. Besides, we provide a comparison of two classes of connectives (primary vs. secondary ones). Our research has shown that these two classes differ in their properties.

[2] http://linghub.ru/ru-rstreebank/

2. Background

In our research, the underlying **discourse structure representation** is the *Rhetorical Structure Theory* [RST; Mann, Thompson 1988]. *It assumes, that a* text is organized into a hierarchical non-projective tree where discourse units (text spans) of smaller size are embedded into bigger ones. Discourse units are connected to each other by rhetorical relations. We concentrate our attention only on asymmetric relations, in which one of the text spans, the Nucleus, carries more important information than the other one, the Satellite, as in (*Peter went home*)$_{nucleus}$ (*because he was tired*)$_{satellite}$ [Mann, Thompson 1988].

As we are dealing with written texts, we consider clauses as elementary discourse units (EDUs), and not prosodic units (as in [Hirschberg, Litman 1993]; [Chafe 1994]; [Kibrik, Podlesskaya 2003]). Structures smaller then a finite clause, such as nominalized constructions or infinitival clauses, can also be treated as EDUs [Carlson, Marcu 2001]; [Schauer 2000]. For example, a preposition can signal causal relations between its dependant expressed via nominalization and the rest part of a clause as in ***Из-за*** (*его позднего возвращения*)… 'due to his late return').

The rhetorical relations quite often are signalled via special lexical clues (discourse connectives). Thus, the **construction of a discourse connectives lexicon** is an essential task. One of the possible approaches is to compile a repository of connectives manually, using standard dictionaries and grammars (e.g. dictionary of connectives for German and English - DiMLex [Stede, Umbach 1998]), for Spanish [Alonso et al. 2002], for French [Roze et al. 2012]. Another way to compile a list of connectives is extracting them from available corpora, e.g. Arabic lexicon [Al-Saif et al. 2010], the list of connectives for Russian [Toldova et al. 2017]. Finally, a list of connectives extracted from existing corpora for a source language can be translated into the target language [Meyer, Webber 2013].

Another relevant task is to settle a set of **annotation features** for **connectives classification**. In [Grote, Stede 1998] authors propose **syntactic** features such as part-of-speech, type of connection it establishes, scope of a connective, linear ordering of the conjuncts, connective position within a text span, **semantic** (semantic relations, polarity, functional ordering of spans) and **pragmatic** ones (discourse relation, presupposition, stylistic features). Some contextual features (occurrence in initial/final sentence or segment, previous/following word, level of embedding etc.) are mentioned in [Alonso et al. 2002]. In the Penn Discourse Treebank (PDTB) approach (cf. international multilingual project TED [Lee et al 2016]), discourse connectives are treated as discourse-level predicates that have two arguments – text spans referring to events or states [Prasad et al. 2007]. We consider this approach while working out our own scheme of connectives annotation.

We also take into account the typology of causal relations signals [Asghar 2016; Chang, Choi 2006]; [Khoo 1998]. According to [Khoo 1998], these are the following types of devices: (1) causal connectives linking two phrases, clauses or sentences (adverbial, prepositional, clause-integrated connectives); (2) causative verbs - transitive verbs that specify the result of an action, event or state, or the influence of some object; (3) resultative constructions; (4) conditional constructions; (5) causative adverbs and adjectives. Only type (1) and type (2) devices are represented in our corpus in sufficient quantity.

Thus, connectives lexicons usually represent morphological, syntactic, semantic and pragmatic description of connectives, the constraints on linear position of discourse units they connect and some other configurational properties.

3. Data

3.1. The current study is based on the **relations annotated in the Ru-RSTreebank** [Pisarevskaya et al. 2017]. The corpus consists of 79 texts, including news, news analytics and popular science (5582 EDUs and 49,840 tokens in total). The text segmentation in the corpus satisfies the principles suggested in [Carlson, Marcu 2001]. Besides EDUs, corresponding to finite clauses, it contains intra-clausal EDUs [cf. Schauer, 2000]. Thus, prepositional phrases, adverbial phrases headed by corresponding connectives (cf. *because of, in spite of*) are treated as separate EDUs.

We limit our investigation only to one type of discourse relations, namely, causal ones. There are 220 examples for the Cause-Effect relation and 110 for the Evidence in our corpus, including both intra- and inter-sentential relations. All these examples are annotated in terms of discourse connectives and their properties.

3.2. The list of **connectives** is manually extracted from the examples. The connectives are divided into two classes, primary vs. secondary connectives, according to their degree of grammaticalization [cf. Rysová and Rysová, 2014, 2015]. As the degree of grammaticalization is a gradable feature rather than a binary one, we rely on the Russian grammar [Shvedova ed. 1980] as a reference source.

Primary connectives are "mainly grammatical (or functional) words which primary function is to connect two units of a text" [Rysová M., Rysová K. 2014], e.g. *из-за* 'because of', *поэтому* 'therefore' etc. We treat the connectives that are enumerated in Russian grammar [Shvedova ed. 1980] as primary connectives. Thus, some of the multi-word prepositions are also treated as primary connective (e.g. *в связи с* 'in connection with').

Secondary connectives, known as Alternative Lexicalization (AltLex) in the Penn Discourse Treebank [Prasad et al. 2010], are not yet fully grammaticalized. These are, primarily, multi-word expressions, e.g. *это привело к тому, что* 'this led to the fact that', *причина этого…* 'the cause is…' etc. These connectives are quite frequent in our corpus. They occur in 46.5% of our examples.

Thus, the class of secondary connectives is of special interest. It constitutes a heterogeneous and open-ended class of elements. Our goal is to single out basic patterns for secondary connectives formation.

The final list of discourse connectives for causal relations consists of 48 connectives (see Fig. 1 for the most frequent of them).

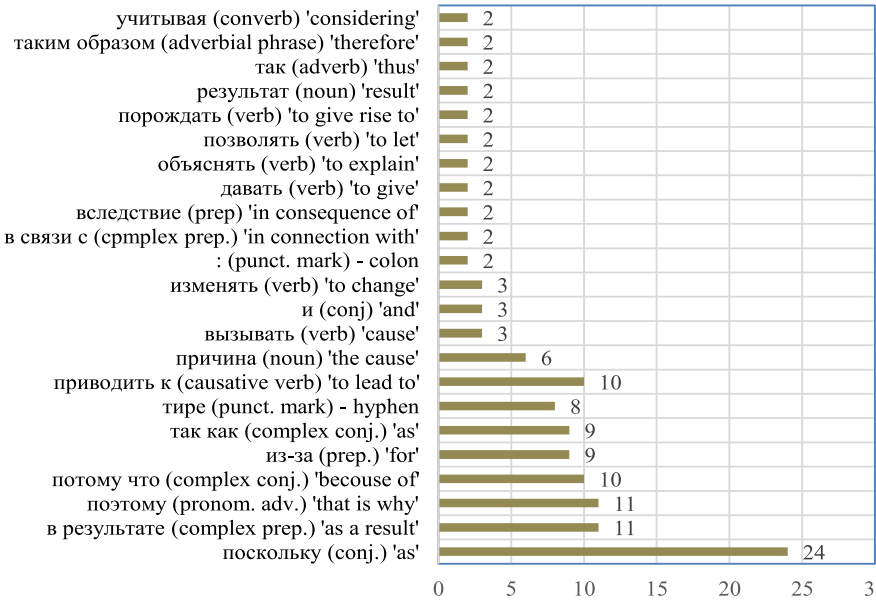


Figure 1. Discourse connectives frequency in Ru-RSTreebank (frequency >1)

3.3. Relying on the essential works devoted to the construction of connective lexicons [Grote, Stede, 1998]; [Alonso et al. 2002]; [Mírovský et al. 2017], we suggest **a scheme** for connectives **annotation** and annotate all the occurrences of connectives in our examples. The scheme includes the properties of connectives (whether it is a multi-word expression (MWU) or not, POS of its core word (for core word see 4.1.); (b) whether a connective is mentioned in the certain resources (see 3.4 for details); (c) the position of a connective in a clause and in a sentence; (d) the properties of the arguments of a connective, such as their complexity, their position and their grammatical features (whether they are headed by a finite verb or by non-finite verb forms). Our final set of features is presented in Table 1.

Table 1. The annotation scheme with an example: the annotation of 'Что **приводит к тому, что'** [that leads to...]

Feature	Values	Example
Type of connective	primary/secondary/NA	secondary
Simple or complex structure	simple/compound	MWU
Listed in the RNC MWU lists	yes/no	no
A causal conjunction listed in RusGram	yes/no	no
Core word in a connective	verb	verb
Position of a connective within EDU	clause initial / clause internal	clause initial

Feature	Values	Example
Position of a connective within a sentence	sentence initial/no	no
Position of a connective wrt. nucleus vs. satellite	nucleus/satellite	nucleus
Connection type wrt sentential boundaries	intra-/inter-sentensial	inter-sentential
Order of Nucleus (wrt. satellite)	1/2	
Occurrence with an anaphoric element	anaphoric element	yes
Satellite complexity	span, sentence, multiclause, clause	clause
Nucleus complexity	span, sentence, multiclause, clause, subclause	clause
Number of clauses in the Nucleus	number	1
Number of clauses in the Satellite	number	1
Gram. features of the Nucleus head	Indicative/converb/participle/nominalization/noun	indicative
Gram features of the Satellite head	Indicative/converb/participle/nominalization/noun	indicative
Anaphoric element in the Nucleus	this/that/what/nd	what
Anaphoric element in the Satellite	this/that/what/nd	nd
Discontinuity	no	no

We follow [Rysová M., Rysová K. 2014] in that we include the anaphoric elements into the connective annotation as in *в следствие этого* 'in consequences of this'. We take into account the demonstrative *это* 'this' (***Это*** *могло стать причиной* 'This could have caused'), wh-relative pronouns as *что* (***Что*** *могло стать причиной*) and expressions with *то* 'that' (cf. *то, что* 'the fact, that'). We also register what text span (Nucleus vs. Satellite) an anaphoric element refers to (see 4.2. for details).

3.4. We compare our results with the theoretical works devoted to expression of causal relations in Russian. We check the resulting list of connectives against the existing resources for functional words and phrases.

There is a detailed survey of causal subordinate conjunctions in RusGram[3] [Pekelis 2014], both simple (*так как*; *поскольку* 'as, since' *etc.*) and complex (*благодаря тому* (,) *что* 'due to'*; в результате того (,) что etc.* 'as a result'*).* This survey concerns inter-clausal relations, though the author mentions that some of the

[3] http://rusgram.ru/%D0%9F%D1%80%D0%B8%D1%87%D0%B8%D0%BD%D0%BD%D1%8B%D0%B5_%D0%BF%D1%80%D0%B8%D0%B4%D0%B0%D1%82%D0%BE%D1%87%D0%BD%D1%8B%D0%B5#12

connectives can signal intra-clausal relations, e.g. *В результате этого* 'As a result of this'. We also have checked our list of connectives against the Russian National Corpus lists of functional MWUs[4].

A list of content words expressing cause-effect relation is given in Ju. Apresyan [Apresyan 2001], such as causal verbs (e.g. *вызывать (болезнь)* 'to cause (a disease)', *внушать (ужас)* 'to excite (a horror)') and nouns (e.g. *основание* 'a ground', *мотив* 'a motive'). The detailed analysis of lexemes expressing "cause" or "purpose" is proposed in [Boguslavskaya, Levontina 2004]. This work is devoted to lexicographic issues. However, both works provide lists of potential content words for 'cause-effect' connectives.

4. Causal connectives in Ru-RSTreebank and their properties

4.1. Patterns for Secondary Discourse Connectives signalling causal relations

As a result of corpus analysis, the basic patterns for secondary connectives formation were distinguished. Our classification is based on the part of speech of the core word. According to [Mírovský et al. 2017], the core word of a connective is "the word that most strongly signals the relation that the whole connective expresses". The types of secondary connectives are presented below.

Constructions containing causative verbs; as causative verbs we treat the verbs whose meanings include a causal element [Asghar 2016] such as verbs of causation (*X позволяет Y* 'X enables Y', *X вызывает Y* 'to produce' etc.), verbs of mental impact [Paducheva 2004]; [Glovinskaya 1993] (c.f. *X можно объяснить Y-ом* 'one can justify X via Y'), motion causation verbs (*X приводит к Y* 'to bring about'), change of state causation verbs (*X изменяет Y* 'X causes the change in Y') and some other (*X порождает Y* 'X gives rise to'):

(1) *[Неудачно остановившаяся машина стала помехой для быстрых кругов многих гонщиков, включая Фернандо Алонсо,] [и* ***это вызвало*** *расследование FIA.]*
'The poorly stopped car became a hindrance to the fast laps of many racers, including Fernando Alonso, and **this evoked** an investigation by the FIA.'

In (1) the connective X *вызывать Y* is in nucleus, Y is a nominalization while X is an anaphoric element *это* 'this', both elements X and Y are located in the same EDU (nucleus), demonstrative *это* substitutes the satellite proposition.

Light verbs constructions, that are structures including a content noun denoting 'cause' or 'effect' (*причина* 'cause', *результат* 'result', *повод* 'matter', *основа* 'basis', *основание* 'basis', *вывод* 'conclusion', *отправная точка* 'starting point', *подтверждение* 'confirmation', *довод* 'argument', *свидетельство* 'evidence' etc.) and a light verb (*являться* 'to be', *становиться* 'to become', *давать* 'to give' etc.):

[4] http://ruscorpora.ru/obgrams.html

(2) *[… повышение цен* ***стало результатом]*** *[удорожания сырья.]*
'… the price increase **was the result** of the rise in price of raw materials.'

In (2) both arguments of the connective are nominalizations.

Complex prepositions, or secondary prepositions in term of the CzeDLex [Rysová, Rysova 2014], that usually are composed of a preposition and a content noun (*в результате X* 'as a result', *вследствие X* 'in consequence of', *в отместку за X* 'in revenge'), excluding prepositions mentioned in [Shvedova ed. 1980]:

(3) *К тому же неприличным был объявлен текст песни «Ich tu dir weh»,*
в результате чего *композицию запретили для исполнения на публике.*
'In addition, the lyrics of the song "Ich tu dir weh" were declared indecent,
as a result of which the composition was banned for performance in public.'

Adverbials: adverbial phrases [Kustova 2017] and converbs; (X *связан с Y* 'concerned with', *X обусловлен Y* 'caused by', *учитывая X* 'taking into account'):

(4) *Предполагают, что синий цвет внешнего кольца* ***обусловлен тем, что*** *оно в дополнении к пыли обладает некоторой примесью мелких частиц водяного льда с поверхности Маба.*
'It is assumed that the blue color of the outer ring **is due to the fact that**
in addition to the dust it has some admixture of small particles of water ice from the surface of the Mab.'

Adverbs *((и) поэтому+CP* '(and) therefore', *(а) потому* '(and) that is why', *не случайно+CP* 'not accidentally'):

Other *constructions with nouns* *(X одна из основ Y* 'one of the bases', *как следствие* 'as a consequence', *причина ADJ:* 'a cause is ADJ'):

(5) *Нынешний норматив — 75% — был введен законом с подачи ЦБ вскоре после «великого дефолта».* ***Причина простая****: финансовый кризис лишил его валютных резервов.*
'Current norm—75% was established by the law at the CB behest soon after the "great default". **The reason (**for that) **is simple**…'

The proportion of different types of connectives, with respect to the core word, is given in Fig. 2:

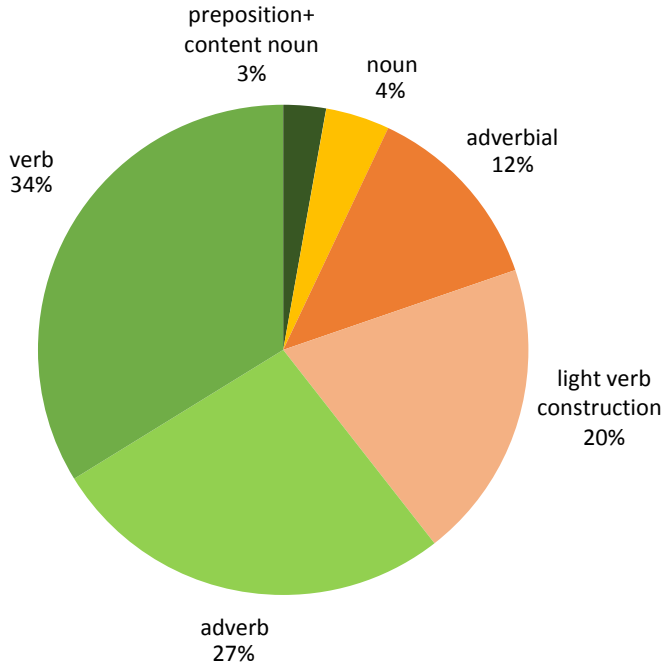


Figure 2. Core words of secondary connectives for 'Cause-Effect' relation (the grammaticalized complex prepositions 'preposition+content noun' are excluded)

4.2. Anaphoric Elements in Multi-Word Connectives

As it has been mentioned, the connectives can contain **anaphoric and cataphoric expressions** such as the demonstrative *это (этот)* 'this' or the relative pronoun *что* 'what':

(6) *В то же время некоторые ведомства, в частности МЧС, готовы экспериментировать … **Благодаря этому** они надеются резко повысить зарплату госслужащим ….* [НКРЯ]
'At the same time some of the departments are ready to carryout experiments… **Due to this** they hope to raise wages of state employees.'

(7) *Многие современные лекарственные препараты включают несколько ингредиентов, **благодаря чему** достигается большая эффективность.*
'Many modern pharmaceuticals include several ingredients, **that's why** (lit. thanks to which) the greater effect is achieved'

Этот is used for inter-sentential connection, *что* - for intra-clausal connection. While the expressions ***благодаря** этому* 'due to this' and ***благодаря** чему* 'lit. due to what' are not fully grammaticalized, the expressions with the same core word

containing *то* 'that' (e.g. ***благодаря*** *тому, что* 'due to the fact that') is a complex conjunction included in dictionaries and grammars. There is also a parallel nominalized construction (as in ***благодаря*** *[его возвращению]*$_{EDU}$ 'due to his return'). This construction is a part of a finite clause. It constitutes an EDU. In this case, the connective can consist only of a core word and can be classified as a primary connective.

To sum up, the core word of some connectives can be combined with all of the mentioned above pronouns as well as with nominalized constructions. Thus, there are four possible constructions with the same meaning.

4.3. Comparison of Primary and Secondary Connectives

We conducted a comparative analysis of two classes of connectives, namely secondary vs. primary ones. The comparison is limited to the 'Cause-Effect' subset (141 example with overt connectives from 155 in total). The aim is to single out the features of these two classes that can help in further connectives extraction and classification.

Firstly, there is a statistically significant difference in the position of two types of connectives with respect to the satellite vs. nucleus opposition (a cause is a satellite and an effect is a nucleus in 'Cause-Effect' relation). Primary connectives are more often located in satellites, while secondary "prefer" nuclei (cf. Table 2, excluding 6 examples where connectives are in both EDUs). Therefore, primary connectives are located in an effect-span more frequently than secondary ones.

Table 2. The position of a connective in nucleus vs. satellite EDU with respect to connective type

connective type / position	nucleus	satellite	sum
primary	22	59	**81**
secondary	37	17	**54**
sum	59	76	**135**
$\chi^2(1) = 20.88, p < .001$ (Yates' correction)			

The order of EDU differs, depending on the class of the connective used to signal the relation between EDUs. The preferable order is "nucleus-satellite" for relations marked with primary connectives and the satellite precedence is preferable with secondary ones (table 3, excluding 1 discontinuous EDU):

Table 3. The position of nucleus EDU in relation: primary vs. secondary connectives

connective type / span order	nucleus precedes	satellite precedes	sum
primary	44	38	**82**
secondary	19	39	**58**
sum	63	77	**140**
$\chi^2(1) = 5.18, p = .023$ (Yates' correction)			

The primary connectives signal inter-clausal relations within a sentence more frequently than secondary ones, while the latter are used to mark inter-sentential relations or they are used for intra-clausal relations when discourse units are expressed via nominalizations:

Table 4. The difference between two classes of connectives wrt of signalling itra- vs. inter-clausal relations

connective type / type of connection	intra-clausal	inter-clausal	intra-sentential	sum
primary	61	11	11	**83**
secondary	26	19	13	**58**
sum	87	30	24	**141**
$\chi^2(2) = 12.34, p = .002$				

There is no statistically significant difference between spans size with secondary and primary markers.

5. Conclusion

In this paper, we present the analysis of discourse connectives in Russian. The analysis is based on detailed examination of 'Cause-Effect' connectives. In our research, we consider the connectives used for signaling the relations between text spans of different size (clause, sentence or bigger). We pay special attention to less grammaticalized constructions for Russian.

As a result, we suggest the list of causal relation connectives based on Ru-RSTreebank (it includes 48 elements), schemes for connectives annotation in the corpus and in the lexicon. All the examples of 'Cause-Effect' relation are annotated according to the corresponding scheme. The basic patterns for non-grammaticalised connectives, including verb and light verb constructions are determined. These patterns can be exploited to expand the list of causal connectives automatically via mining them in an unannotated corpus.

Besides, we provide the comparison of two classes of connectives. Our data show the statistically significant difference between primary vs. secondary connectives with respect to the nucleus/satellite position, intra- vs. inter-sentential relations and some others. The features that exhibit such a difference could be taken into consideration while developing machine learning technique for rhetorical relations extraction.

References

1. *Al-Saif A., Markert K.* (2010), The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. In LREC 2010 Proceedings, pp. 2046–2053.
2. *Alonso L., Castellón I., Gibert K., Padró L.* (2002), An empirical approach to discourse markers by clustering. In Topics in Artificial Intelligence, Springer, Berlin, Heidelberg, pp. 173–183.
3. *Apresyan YU. D.* (2001), System-forming meanings 'to know' and 'to consider' in Russian [Sistemoobrazuyushchiye smysly «znat'» i «schitat'» v russkom yazyke], Russian Language and Linguistic Theory [Russkiy yazyk v nauchnom osveshchenii], 1, pp. 5–26.
4. *Asghar N.* (2016), Automatic Extraction of Causal Relations from Natural Language Texts: A Comprehensive Survey, arXiv preprint arXiv:1605.07895.
5. *Boguslavskaya, O. YU., Levontina, I. B.* (2004). Meanings 'cause' and 'purpose' in natural language [Smysly 'prichina' i 'tsel' v yestestvennom yazyke], Topics in the study of language [Voprosy yazykoznaniya], (2), pp. 68–88.
6. *Carlson L., Marcu D.* (2001), Discourse tagging reference manual. ISI Technical Report ISI-TR-545, 54, 56.
7. *Chafe, W.* (1994), Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing, University of Chicago Press.
8. *Chang D. S., Choi K. S.* (2006), Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities , Information processing & management, Vol. 42, №. 3, pp. 662–678.
9. *Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., ... & Schlaefer, N.* (2010), Building Watson: An overview of the DeepQA project. AI magazine, 31(3), 59–79.
10. *Galitsky B., Ilvovsky D. and Kuznetsov S. O.* (2018), Detecting Logical Argumentation in Text via Communicative Discourse Tree. To appear in Journal of Experimental and Theoretical Artificial Intelligence. JETAI 2018.
11. *Glovinskaya, M. Y.* (1993), Russian Speech Acts with the Meaning of Mental Pressure./ Logical analysis of language. Mental actions. Moscow: Science. [Russkie rechevihe akty so znacheniem mentaljnogo vozdeyjstviya / Logicheskiyj analiz yazihka. Mentaljnihe deyjstviya.]
12. *Grote B., Stede M.* (1998), Discourse marker choice in sentence planning, Proceedings of the 9th International Workshop on Natural Language Generation, Niagara-on-the-Lake, Canada, August 1998.
13. *Hirschberg, J., Litman, D.* (1993), Empirical studies on the disambiguation of cue phrases, Computational linguistics, 19(3), pp. 501–530.
14. *Khoo C. S. G. et al.* (1998), Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing, /Literary and Linguistic Computing, Vol. 13, №. 4. pp. 177–186.
15. *Kibrik, A. A., Podlesskaya, V. I.* (2003), Towards building corpora of oral Russian speech: principles of transcription [K sozdaniyu korpusov ustnoy russkoy rechi: printsipy transkribirovaniya], Scientific and Technical Information Processing (series 2) Nauchno-tekhnicheskaya informatsiya (seriya 2), 6, pp. 5–11.

16. *Knott A.* (1996), A data-driven methodology for motivating a set of coherence relations, Thesis (doctoral), University of Edinburgh.
17. *Kustova, G. I.* (2017), The Types of Adverbials [Tipy konstruktsiy s adverbialami], Computational Linguistics and Intellectual Technologies: Proc. of the Int. Conf." Dialogue, Vol. 2, pp.234–249.
18. *Lee, A., Prasad, R., Webber, B., & Joshi, A. K.* (2016), Annotating Discourse Relations with The PDTB Annotator. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, pp. 121–125).
19. *Louis, A., Joshi, A., & Nenkova, A.* (2010), Discourse indicators for content selection in summarization. In Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue (pp. 147–156), Association for Computational Linguistics.
20. *Mann W. C., Thompson S. A.* (1988), Rhetorical Structure Theory: Toward a Functional Theory of Text Organization, Text 8, 3, pp. 243–281.
21. *Meyer T., Webber B. Implicitation of discourse connectives in* (machine) translation (2013), Proceedings of the Workshop on Discourse in Machine Translation, pp. 19–26.
22. *Mírovský, J., Synková, P., Rysová, M., & Poláková, L.* (2017), CzeDLex–A Lexicon of Czech Discourse Connectives. The Prague Bulletin of Mathematical Linguistics, 109(1), pp. 6191.
23. *Paducheva, E. V.* (2004), Dynamic models in the semantics of vocabulary [Dynamicheskiye modeli v yazyke]. M.: Languages of Slavonic Culture.
24. *Pekelis, O. Ye.* (2014), Causal subordinate clauses [Prichinnyye pridatochnyye], Materials for the project of Russian grammar corpus description [Materialy dlya proyekta korpusnogo opisaniya russkoy grammatiki], avaliable at: http://rusgram.ru. As a manuscript [Na pravakh rukopisi], M, 2015.
25. *Pisarevskaya D. et al.* (2017), Towards building a discourse-annotated corpus of Russian, Computational Linguistics and Intellectual Technologies: Proc. of the Int. Conf.» Dialogue, Vol. 1, pp. 194204.
26. *Prasad R. et al.* (2007), The penn discourse treebank 2.0 annotation manual.
27. *Prasad R., Joshi A., Webber B.* (2010), Realization of discourse relations by other means: alternative lexicalizations, Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, pp. 1023–1031.
28. *Roze C., Danlos L., Muller P.* (2012), LEXCONN: a French lexicon of discourse connectives, Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics, №. 10.
29. *Rysová M., Rysova K.* (2014), The Centre and Periphery of Discourse Connectives, PACLIC, pp. 452–459.
30. *Schauer H.* (2000), From elementary discourse units to complex ones, Proceedings of the 1st SIGdial workshop on Discourse and dialogue-Volume 10, Association for Computational Linguistics, pp. 46–55.
31. *Shvedova, N. YU. ed.* (1980), Russian grammar [Russkaya grammatika]. V dvukh tomakh. AN SSSR Institut russkogo jazyka. — M.: Nauka, 1980.

32. *Stede M., Umbach C.* (1998), DiMLex: A lexicon of discourse markers for text generation and understanding, Proceedings of the 17th international conference on Computational linguistics-Volume 2, Association for Computational Linguistics, pp. 1238–1242.
33. *Taboada, M., Voll, K., & Brooke, J.* (2008), Extracting sentiment as a function of discourse structure and topicality. Simon Fraser Univeristy School of Computing Science Technical Report.
34. *Toldova S. et al.* (2017), Rhetorical relation markers in Russian RST Treebank, Proceedings of the 6th Workshop Recent Advances in RST and Related Formalisms, Santiago de Compostela, Spain, September 4 2017, Association for Computational Linguistics, pp. 29—33.

СИНТАКСИС ПРЕДЛОГООБРАЗНЫХ НАРЕЧИЙ: НЕКОТОРЫЕ СЛОЖНЫЕ СЛУЧАИ[1]

Урысон Е. В. (uryson@gmail.com)
ИРЯ РАН, Москва

SYNTAX OF PREPOSITIONAL ADVERBS: SOME DIFFICULT CASES

Uryson E. V. (uryson@gmail.com)
Russian Language Institute (RAS), Moscow, Russia

The subject of this paper are Russian so called adverbial prepositions; cf. vokrug (kostra) ‘around smth.’, daleko ot (doma) ‘far from smth.’, etc. By definition, an adverbial preposition either coincides with an adverb (cf. vokrug) or contains an adverb and a preposition (cf. daleko ot). As I have demonstrated in my previous works, an adverbial preposition and the underlying adverb have the same meaning, the only difference between them being in 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’. From the modern point of view, syntactic distinction is insufficient for interpreting such cases as different words (or different meanings of a word). So, an adverbial preposition and the underlying adverb should be interpreted as the same meaning of a given word. I argue that this word is an adverb (or a prepositional adverb). This paper deals with syntax of these adverbs. Such adverbs have one or more semantic actants, at least one of them being expressed by a noun or a prepositional group. The problem is that in some cases it is not clear whether the prepositional group is governed by the adverb or by the verb governing this adverb (thus the adverb and the prepositional group are co-governed by the verb). A criterion of adverb vs. verb governing of such groups is discussed. Two Russian adverbs zadolgo ‘for a long time before smth.’ and nezadolgo ‘for a long time before smth.’ are described from this point of view.

Keywords: Adverbial preposition; adverb; semantic actant; syntactic actant; obligatory expression of an actant; offset of a complement

[1] Работа выполнена при финансовой поддержке РФФИ (проект № 16-04-00302).

0. Объект исследования и постановка проблемы

Объект исследования — русские лексические единицы, которые академическая грамматика [РГ-70; РГ-80] трактует как наречные предлоги. Ср. *вокруг (дома), далеко от (работы), рядом с (гостиницей)*. По определению, наречный предлог либо формально совпадает с каким-либо наречием, либо в нем выделяется компонент, формально совпадающий с наречием. Так, наречный предлог *вокруг* совпадает с наречием *вокруг*; ср.

(1) *Вокруг росли деревья* (наречие).

(2) *Вокруг дома росли деревья* (наречный предлог).

В наречном предлоге *далеко от* есть компонент *далеко*, формально совпадающий с наречием; ср.

(3) *Он живет далеко* (наречие).

(4) *Он живет далеко от работы* (наречный предлог).

В предыдущих работах [Урысон 2017; 2014] было продемонстрировано, что наречие и формально совпадающий с ним (возможно частично) наречный предлог имеют одну и ту же семантику. В частности, они имеют один и тот же набор семантических актантов: так, *вокруг* и *далеко* в обоих контекстах имеют семантический актант ‘пространственный ориентир’. Различие между контекстами типа (1) и (2) или (3) и (4) сводится к способу оформления данного семантического актанта. В случаях типа (2) и (4) этот актант оформляется падежной формой существительного или предложно-падежной группой, которая синтаксически подчиняется обсуждаемой единице: *вокруг → дома, далеко → от работы*. Что касается контекстов типа (1) и (3), то здесь тот же семантический актант выражается каким-либо словом обычно в предтексте, причем способ оформления этого актанта не может быть описан четкими морфологическими или морфосинтаксическими правилами. Для обнаружения в тексте подобного актанта требуются правила иного рода (вероятно, они подобны правилам поиска антецедента местоимения).

С точки зрения академической грамматики слово *вокруг* в контекстах (1) и (2) представлено в разных значениях, или (в терминологии московской семантической школы) в контекстах (1) и (2) представлены разные лексемы слова *вокруг*. Различаются эти лексемы своей частеречной принадлежностью. Аналогичным образом, в (3) и (4) тоже представлены разные лексемы слова *далеко*, различающиеся только частью речи.

С точки зрения современного подхода к описанию слов, в случаях типа (1) и (2) или (3) и (4) представлена одна и та же лексема (*вокруг* или *далеко* соответственно). Ее семантический актант ‘пространственный ориентир’ может быть выражен двумя способами: (а) словом, синтаксически не связанным непосредственно с данным словом; (б) формой определенного падежа существительного (или определенной предложно-падежной группой). Эта информация помещается в синтаксической зоне словарной статьи лексемы. Остается грамматический вопрос: к какой части речи относится данная лексема?

Имеется всего два варианта, альтернативных академическому описанию: отнести лексему *вокруг* или *далеко от* во всех случаях к предлогам или во всех случаях — к наречиям. Неудовлетворительность первого подхода очевидна. Во-первых, придется считать, что предлог одинаково свободно может выступать как с зависимой формой, так и без нее, ср. равно употребительные и нейтральные контексты типа (1) и (2), (3) и (4); ясно, что такое допущение размывает само определение предлога. Во-вторых, придется признать, что предлог типа *далеко от* теряет свой второй компонент (*от*) в случаях, когда выступает без зависимого слова; соответствующее описание будет, очевидно, громоздким и грамматически нелогичным.

Поэтому в работах [Урысон 2017; 2014] предлагается считать, что как в контексте типа (1) или (3), так и в контексте типа (2) или (4) обсуждаемые лексемы *вокруг* и *далеко (от)* являются наречиями. Тем самым, мы признаем, что наречие может управлять существительным (его падежной формой) или предложно-падежной группой. Может показаться, что такое допущение противоречит принятому определению наречия. Однако это не так: в классе наречий выделяются предикативные наречия, а они способны управлять подобно глаголу. Ср. *жаль кого-л. (Мне жаль его); стыдно за кого/что-л. (Было стыдно за жену, за этот поступок); страшно за кого-л. (Ей страшно за детей); можно (Ему можно гулять); темно, сыро, душно (Там темно и сыро, Мне здесь душно)* и т. п. Обсуждаемые наречия, хотя и не являются предикативными, но тоже способны управлять.

Итак, в классе наречий выделяются управляющие наречия, которые в свою очередь делятся на два класса: предикативные наречия и те, которые составляют объект нашего исследования (наречные предлоги академической грамматики). Наречия последнего класса будем называть предлогообразными.

Подчеркнем основное отличие предлогообразного наречия от предлога: любой предлог обязательно управляет падежной формой: высказывание с предлогом, не имеющим такой зависимой формы, практически невозможно или воспринимается как очень экспрессивное (ср. *Да не НА столе, а ПОД, ПОД!)*. Наречие, даже если оно способно управлять падежной формой (ср. *вокруг дома*), может употребляться и без нее (ср. *Мы стояли вокруг и смотрели*). Кроме того, предлогообразное наречие может управлять не падежной формой, а предложно-падежной группой (ср. *далеко от музея*).

Основная синтаксическая проблема формулируется так. В синтаксической структуре предложения предлогообразное наречие, очевидно, зависит от глагола. Предложно-падежная группа может зависеть от этого наречия (наречное управление), а может тоже подчиняться глаголу. В последнем случае имеет место соподчинение глаголу наречия и данной группы. В некоторых случаях неясно, чему синтаксически подчиняется данная группа — наречию или глаголу. Ср.

(5) *Единственная тусклая лампочка висит очень высоко над столом*.

Это предложение как будто допускает два анализа:

(6) а *висит → высоко, висит → (над столом)* [соподчинение наречия и предложно-падежной группы глаголу-сказуемому].
б *висит → высоко → (над столом)* [наречное управление].

Цель работы — уточнить критерий определения наречного управления. Заметим, что синтаксическая зависимость падежной формы в реальном высказывании устанавливается однозначно (ср. *стоять → вокруг → елки*). Поэтому в дальнейшем нас будут интересовать только предложно-падежные группы в контексте наречия.

Последовательное различение наречного управления и соподчинения наречия и предложно-падежной группы некоему третьему слову могли бы лечь в основу критерия выделения составных наречных предлогов. Однако, насколько нам известно, в рамках академического подхода данная проблематика не обсуждается.

Дальнейшее изложение строится по следующему плану. В разделе 1 на материале некоторых пространственных наречий обсуждается критерий наречного управления. В **Разделе 2** на примере некоторых временны́х наречий рассматривается случай обязательного выражения семантического актанта наречия и предлагается уточнение обсуждаемого критерия.

1. Наречие и предложно-падежная форма: управление или соподчинение?

Для дальнейшего нам понадобится критерий определения главного слова в словосочетании, сформулированный в работах [Курилович 1962; Холодович 1979]:

Если удалить главное слово словосочетания, то нарушится грамматическая правильность предложения.

(7) *В вазе стоит белая ← роза*. **В вазе стоит белая*. vs. *В вазе стоит роза*.

Следовательно, в словосочетании типа *белая роза* синтаксически главным является существительное *роза*.

Однако наша задача состоит не в определении синтаксически главного слова в словосочетании, а в том, чтобы понять, связаны ли данные слова непосредственной синтаксической зависимостью или они соподчинены третьему слову в предложении. Для ответа на этот вопрос воспользуемся одним очевидным следствием из критерия Куриловича — Холодовича:

Если два слова L1 и L2 в предложении соподчинены некоторому третьему слову и, следовательно, ни L1 не подчиняет себе L2, ни L2 не подчиняет себе L1, то как L1, так и L2 можно удалить из предложения вместе с их зависимыми, и при этом грамматическая правильность предложения не нарушится[2].

Рассмотрим с точки зрения этого критерия примеры (5) и (7):

(5) *Единственная тусклая лампочка висит очень высоко над столом*.

(8) *Единственная тусклая лампочка висит очень далеко от стола*.

[2] Ниже этот критерий будет уточнен. Предлагаемая формулировка не исключает возможности удалить из предложения и L1, и L2 одновременно, однако этот случай не рассматривается.

В первом приближении, *высоко над А2* ‘на большом расстоянии по вертикали вверх от объекта А2’. Тем самым, наречие *высоко* имеет семантический актант ‘пространственный ориентир’. В примере (5) этот семантический актант выражен группой *над столом*. Значит ли это, что наречие *высоко* синтаксически подчиняет себе группу *над столом*? В соответствии с нашим критерием — нет. Действительно, из предложения (5) можно удалить как наречие *высоко* (вместе с его зависимым *очень*), так и группу *над столом*, и грамматическая правильность предложения не нарушится. Ср.

(9) а *Единственная тусклая лампочка висит над столом.*
б *Единственная тусклая лампочка висит очень высоко.*

Следовательно, и наречие *высоко*, и группа *над столом* синтаксически зависят от третьего слова в предложении (глагола-сказуемого *висит*). Тем самым, выражение семантического актанта ‘пространственный ориентир’ наречия *высоко* подчиняется не этому наречию, а глаголу, от которого зависит само наречие.

Пример (8) устроен иначе. Семантический актант ‘пространственный ориентир’ наречия *далеко* выражен группой *от стола*, которая синтаксически подчиняется наречию. Действительно, если удалить из (8) наречие *далеко* (вместе с его зависимым *очень*), то нарушится грамматическая правильность предложения. Ср.

(10) **Единственная тусклая лампочка висит от стола.*

Между тем, группа *от стола* может быть удалена из предложения (8) без нарушения его грамматической правильности. Ср.

(11) *Единственная тусклая лампочка висит очень далеко.*

Заметим, что семантический актант ‘ориентир’ наречий *высоко* и *далеко* может быть не выражен, ср. (9б), (11). В этом случае подразумевается, что ориентируемый объект находится далеко от наблюдателя (который может совпадать с говорящим).

Итак, и наречие *высоко*, и наречие *далеко* имеют семантический актант ‘пространственный ориентир’. Выражается он у каждого наречия по-своему: у наречия *высоко* — группой *над кем-л./чем-л.*, а у наречия *далеко* — группой *от кого-л./чего-л.* Наречие *далеко* синтаксически подчиняет себе эту группу. Наречие *высоко* синтаксически не подчиняет себе выражение своего семантического актанта: и наречие, и соответствующая предложно-падежная группа синтаксически соподчиняются глаголу (как два обстоятельства места).

Однако и наречие *далеко* не всегда подчиняет себе выражение своего актанта ‘ориентир’. Рассмотрим пример:

(12) *Отшельник жил далеко за рекой.*

Предлог *за* (в данном значении) указывает на расположение объекта относительно двух ориентиров. Первый ориентир — наблюдатель, а второй ориентир — это объект, обозначенный существительным, управляемым предлогом

за (в данном случае это река). Действительно, в (12) местонахождение отшельника описано через отсылку к реке, но при этом еще предполагается, что наблюдатель находится на другой ее стороне.

Эта семантика предлога *за* сохраняется и в контексте наречия *далеко*. Очевидно, однако, что группу *за рекой* в (12) можно интерпретировать как выражение семантического актанта ‘ориентир’ наречия *далеко* (с оговоркой о наблюдателе), ср. *далеко от реки*. Тем не менее, данная группа синтаксически не зависит от наречия *далеко*: как наречие, так и данную группу можно опустить, и грамматическая правильность предложения не нарушится. Ср.

(13) *Отшельник жил далеко.*

(14) *Отшельник жил за рекой.*

Обратим также внимание на то, что предлог *за*, как и другие пространственные предлоги в русском языке, имеет два значение: локативное, ср. *жить за рекой*, и направительное, ср. *уйти за реку*. Первая лексема управляет творительным падежом, а вторая — винительным. Выбор лексемы предлога *за* и, следовательно, падеж существительного, диктуется глаголом. Контекст наречия *далеко* не меняет дела, ср. *Отшельник жил далеко за рекой — Отшельник ушел далеко за реку*. Поэтому естественно считать, что предложно-падежная форма синтаксически зависит от глагола, а не от наречия *далеко*. Таким образом, в (12) наречие *далеко* и предложно-падежная форма *за рекой* синтаксически соподчиняются глаголу.

Перед нами обязательное переподчинение семантического актанта лексемы (*далеко*) другой лексеме (сказуемому) в предложении. Заметим, что переподчинение (не всегда обязательное) семантического актанта одной лексемы другой лексеме в предложении — достаточно распространенное явление в языке. Оно подробно описано на глагольном материале как смещение дополнения. Обзор таких случаев дан в работе [Апресян 2010: 121–122]. Приведем некоторые примеры: *крепко сжимать руки бандита — крепко сжимать бандиту руки*; *обработать его раны — обработать ему раны* [Там же].

Из-за такого переподчинения семантического актанта рассмотренные нами примеры синтаксически не отличаются от случаев, когда соподчиненная наречию предложно-падежная группа не выражает никакого его семантического актанта. Ср.

(15) *Поселок стоял далеко в лесу <на склоне горы>.*

(16) *Орел парит высоко под облаками.*

Отметим одну общую черту всех приведенных контекстов: наречие как бы срастается с последующей предложно-падежной формой, образуя единую группу, просодически оформляемую как синтаксически законченную часть высказывания. Ср. возможные ответы на вопрос *Где? — Далеко от стола, Далеко за рекой, Далеко в лесу, Далеко на склоне горы, Высоко над столом, Высоко под облаками* т. п. В позиции предложно-падежной формы может находиться и другое наречие, ср. *далеко впереди <позади>, высоко слева <справа>*.

Тогда наречие *далеко* (или *высоко*) «прилипает» к нему. Казалось бы, это аргумент в пользу того, что наречие и предложно-падежная форма образуют единство, так что предложно-падежная форма синтаксически зависит от наречия, а не от глагола.

Но подобное единство образуют и другие соподчиненные обстоятельства, ср. *Вчера вечером (я была в театре), В среду 11 апреля (нас ждут в гости), В Москве на Тверской (опять была стрельба)* и т. п. Такие соподчиненные обстоятельства располагаются контактно друг к другу и просодически оформляются как словосочетание. Тем не менее не возникает сомнений в том, что оба обстоятельства синтаксически подчиняются сказуемому.

По-видимому, «срастание» таких соподчиненных обстоятельств в единую группу можно объяснить семантическими причинами. Так, в случае типа *далеко в лесу, высоко над столом* предложно-падежная группа выражает семантический актант наречия. Подобным образом устроено сочетание *вчера вечером*: слово *вечер* (и *вечером*) обозначает определенную часть дня, ср. *субботний вечер, вечером того дня*, и, следовательно, имеет семантический актант со значением ‘какой день’ («вечер какого дня»). Если этот семантический актант не выражен, то подразумевается данный день или день, о котором идет речь. В случае *вчера вечером* слово *вчера* выражает семантический актант слова *вечер*, ср. *вечер вчерашнего дня*. Что касается случаев типа *в среду 11 апреля, в Москве на Тверской*, то здесь, как кажется, ни одно слово сочетания не является семантическим актантом другого. Однако с точки зрения наших общих знаний, улица обычно находится в каком-то населенном пункте; а локализуя событие во времени, часто указывают и день недели, и дату. Таким образом, слова в подобных сочетаниях связаны с точки зрения организации наших знаний об окружающем мире. Однако семантическая (или логическая) структура этих сочетаний не влияет на их синтаксическое представление.

В синтаксической зоне словарной статьи пространственного наречия *далеко* требуется указать, что ее семантический актант ‘пространственный ориентир’ может быть выражен группой *от*+РОД, синтаксически подчиняющейся наречию. В других случаях выражение этого актанта синтаксически не связано с наречием непосредственно.

Аналогичным образом устроен еще целый ряд пространственных наречий, имеющих семантический актант ‘ориентир’. Так, наречия *вдали, вдалеке, невдалеке, поблизости, справа, слева* управляют группой *от кого/чего-л.*, обозначающей данный актант. Ср. *Он стоял вдали <поблизости> от них — *Он стоял от них; Вдалеке <невдалеке> от города виднеется монастырь — *От города виднеется монастырь; Справа <слева> от дороги — ни одного огонька — *От дороги — ни одного огонька*.

Наречие *близко*, которое тоже имеет семантический актант ‘ориентир’, обладает вариативным управлением: этот актант может выражаться группой *от*+РОД или *к*+ДАТ, синтаксически подчиняющейся наречию. Ср. *сидеть близко от сцены — сидеть близко к сцене,* при невозможности **сидеть от сцене, *сидеть к сцене*.

Наречие *вплотную* тоже имеет семантический актант ‘ориентир’ и управляет группой *к*+ДАТ, выражающей этот актант. Ср. *Стол стоит вплотную к стене* (при невозможности **Стол стоит к стене*).

Однако даже такое выражение семантического актанта наречия может синтаксически подчиняться глаголу. Примеры.

(17) *Стол поставили далеко от окна*.

(18) *Стол отодвинули далеко от окна*.

В соответствии с нашим критерием, в (17) группа *от окна* синтаксически подчиняется наречию *далеко*: если опустить в (17) наречие, то получится аграмматичный результат, ср. **Стол поставили от окна*. Однако в (18) можно опустить как наречие *далеко*, так и группу *от окна*, и в обоих случаях предложение будет правильным; ср. *Стол отодвинули от окна*, *Стол отодвинули далеко*. Следовательно, в (18) наречие и данная группа соподчинены глаголу. Тем самым, в ситуации выбора, когда обсуждаемую группу можно интерпретировать как зависящую от глагола или от наречия, предпочтение отдается глаголу (другое решение будет противоречить критерию Куриловича — Холодовича).

Подобная ситуация выбора возникает в контексте тех глаголов, которые сами могут управлять обсуждаемой предложно-падежной группой.

Еще раз заметим, что соподчиненные обстоятельства, из которых к тому же одно является семантическим актантом другого, как бы срастаются в единую цепочку. В результате вся эта цепочка в каком-то смысле функционирует как единое целое и в результате может относиться и к такому глаголу, который сам не может управлять данной предложно-падежной группой. В последнем случае предложно-падежная группа из этой цепочки синтаксически «переподчиняется» наречию.

2. Обязательное выражение семантического актанта наречия: *задолго (до праздника), незадолго (до войны)*

Рассмотрим семантическую организацию слов *задолго* и *незадолго*. Ср. *прийти (A1) задолго до звонка (A2), уйти (A1) незадолго до отхода поезда (A2)*. Данные слова предполагают две ситуации A1 и A2, причем указывают, что A1 имела место раньше A2 и отделена от нее определенным промежутком времени. Ситуации A1 и A2 — это семантические актанты данных наречий: A2 — точка отсчета, относительно которой определяется время ситуации A1.

Актанты A1 и A2 могут быть событиями, и тогда событие A2 (ср. ‘звонок’, ‘отход поезда’) служит временным ориентиром события A1. Если же A2 является не событием, а длящейся ситуацией, то временным ориентиром служит ее начало, ср. *незадолго до войны* ‘до начала войны’. В разговорной речи событие, являющееся временным ориентиром, может пониматься из контекста и обозначаться не отдельным предикатом, а (метонимически) его главным участником. Ср. *Он пришел незадолго до Ивана* ‘Он пришел незадолго до прихода Ивана’.

Семантический актант А2 ‘временной ориентир’ данных наречий выражается предложно-падежным сочетанием *до* + РОД (см. примеры выше) или *перед* + ТВОР. Ср.

(19) *Задолго перед тем, как появились эти опытные исследования, Фитцджеральд уже приложил теоретические исследования Максвелла к объяснению отступлений от ньютоновского закона всемирного тяготения в случае движения комет*. [П. Н. Лебедев. Физические причины, обусловливающие отступления от гравитационного закона Ньютона (1902)].

(20) *Незадолго перед роспуском первой Думы, Азеф организует покушение против Столыпина.* [Б. В. Савинков (В. Ропшин). Воспоминания террориста (1909)].

На первый взгляд, кажется очевидным, что предложно-падежная группа, выражающая семантический актант ‘временной ориентир’ наречия *задолго* или *незадолго*, синтаксически подчиняется этому наречию. Применим, однако, к данным словам критерий наречного управления. Ср.

(21) *Он пришел задолго до начала лекции*.

(22) *Они познакомились незадолго до войны*.

В обоих примерах обсуждаемое наречие можно опустить без нарушения грамматической правильности предложения:

(23) *Он пришел до начала лекции*.

(24) *Они познакомились до войны*.

Значит, наречие *задолго* (или *незадолго*) не управляет предложно-падежной группой *до чего-л./кого-л.* При этом саму эту группу удалить из предложения нельзя: для наречий *задолго* и в особенно *незадолго* употребление без предложной группы нехарактерно или даже недопустимо. Ср.

(25) ??*Он пришел задолго.*

(26) **Они познакомились незадолго*.

Правда, в НКРЯ можно найти подобное употребление данных наречий. При этом *задолго* и *незадолго* ведут себя в этом отношении по-разному. Для *незадолго* употребление без предложной группы нехарактерно или даже почти недопустимо, а *задолго* употребляется так чаще и свободнее. (Возможно, это связано с тем, что антоним, указывающий на больший полюс шкалы, вообще имеет более развитую синтактику.)

Употребление *незадолго* без предложной группы представлено в НКРЯ считанными случаями: в период 1990–2018 гг. это 5 примеров из 1644, т. е. 0,3 %. При этом всего 1 из них — это пример (27) — не носит разговорного характера. Ср.

(27) *Ваксон остановился на задах колхозного клуба, где, казалось, незадолго прошел Мамай*. [В. Аксенов. Таинственная страсть (2007)].

(28) *Вспомнила мой разговор с А.И. незадолго: «И я, если окажусь там, переменюсь?» — «И вы». И так и случилось*. [Л. Чуковская. Александр Солженицын (1962–1995)].

Для слова *задолго* употребление без предложной группы более характерно (25 вхождений из 1448, т. е. 1,7%, в период 1990–2018 гг.) и встречается во вполне обработанных литературных текстах. Ср.

(29) *Готовились к юбилею [Сталина] задолго — всей страной, всем миром*. [Мария Чегодаева. Соцреализм: Мифы и реальность (2003)].

(30) *Программы следующих фестивалей становятся известными задолго, поэтому многие любители музыки могут планировать свои поездки на фестиваль в Туре и выбирать концерты тех или иных, особенно любимых ими исполнителей*. [И. К. Архипова. Музыка жизни (1996)].

(31) *Тютчеву предстоит увидеть в реальности то, что он задолго увидел в своем воображении*. [Лев Аннинский. Бессильный ясновидец // «Дружба народов», 2003].

(32) *Тем августом они собирались всей семьей погостить у друзей в Ленинграде. Билеты были куплены задолго — двенадцатилетний сын и сама Нина давно мечтали об этой поездке*. [Дина Рубина. Несколько торопливых слов любви (2001) // «Новый Мир», 2003].

Не исключено, что употребление *задолго* без предложной группы в современном языке может быть квалифицировано как нарративное или не очень тщательное. Употребление *незадолго* без такой группы в современном языке, по-видимому, свойственно нетщательной речи или является устаревшим.

ПРИМЕЧАНИЕ. В период 1800–1875 гг. употреблений *незадолго* без управляемой группы существенно больше, чем в период 1990–2018 гг.: 22 случая на 899 вхождений, т. е. 2,8%. Ср.

(33) *В избе было темно, хоть глаз выколи; острый запах дыма свидетельствовал, что лучина незадолго угасла*. [Д. В. Григорович. Бобыль (1847)].

(34) *Дьячок Никандр, незадолго прибывший из Славяно-греко-латинской академии, обучал Закону Божию, чтению книг по старому и новому письму и церковному пению*. [А. О. Корнилович. Андрей Безыменный (Старинная повесть) (1832)].

(35) *И дети, которые незадолго играли шумно и весело, теперь утомились*. [Н. Г. Помяловский. Молотов (1861)].

В тот же период 1800–1875 гг. употреблений *задолго* без управляемой группы приблизительно столько же, сколько в период 1990–2018 гг.: 9 вхождений из 168, т. е. приблизительно 2%. Ср.

(36) *Добрая слава задолго предлетела ему в Дагестан.*
[А. А. Бестужев-Марлинский. Письма из Дагестана (1831)].

(37) *Задолго я обещал Гаевскому провести этот вечер у Никитенки, сам писал ему о том вчера и забыл свое обещание, так что другие мне о нем напомнили.* [А. В. Дружинин. Дневник (1845)].

(38) *Великий день девятнадцатого февраля мы встретили восторженно и задолго еще начали осушать в честь его тосты.*
[Ф. М. Достоевский. Бесы (1871–1872)].

Как бы то ни было, в полном описании слов *задолго* и *незадолго* требуется учесть и нейтральные случаи, когда *задолго* и *незадолго* практически требуют предложно-падежной группы, и случаи маркированные, когда эти слова употребляются без такой группы. Опишем сначала нейтральное употребление.

Вернемся к примерам (23)–(24) и (25)–(26). Поскольку из нейтрального высказывания можно удалить наречие *задолго* или *незадолго* (без нарушения грамматической правильности предложения), но нельзя удалить предложно-падежную группу, обозначающую семантический актант ‘временной ориентир’ данного наречия, то, казалось бы, следует считать, что наречие *задолго* (или *незадолго*) синтаксически подчиняется этой предложно-падежной группе, или, более точно — предлогу:

(39) а. *задолго ← до → войны; незадолго ← до → спектакля.*
б. *задолго ← перед → войной; незадолго ← перед → спектаклем.*

Такое описание как будто соответствует критерию Куриловича — Холодовича и, следовательно, вполне имеет право на существование. Тем не менее, оно вызывает большие возражения. Действительно, предлог, вообще говоря, не подчиняет себе никаких слов, кроме управляемой падежной формы. Исключением являются некоторые частицы, ср.

(40) *остаться совсем ← без денег*; *расположиться прямо ← у воды*; *непосредственно ← перед спектаклем.*

Таким образом, структура типа (39) требует большого количества оговорок: в ней допустимы лишь временные предлоги *до* и *перед*, а из знаменательных слов они могут подчинять только наречия *задолго* и *незадолго*.

Более общее решение, не требующее введения уникальной структуры типа (39), состоит в следующем. И наречие *задолго* или *незадолго*, и предложно-падежная группой *до* + РОД или *перед* + ТВОР подчиняются глаголу, т. е. здесь имеет место обязательное синтаксическое переподчинение глаголу семантического актанта наречия. Но при этом сам этот семантический актант тоже должен быть обязательно выражен.

Обязательное переподчинение наречного семантического актанта глаголу обсуждалось выше — это вполне обычная ситуация. Обязательного выражения семантического актанта тоже требует целый ряд предикатов. Ср. глаголы местонахождения *находиться (где-л.), водиться (где-л.), деться куда-л.* и т. п. Все они требуют выражения семантического актанта ‘место’, который оформляется как

синтаксический актант глагола. Нормально: *Эта улица находится в Западном округе* при недопустимости: **Эта улица находится*. Нормально *Слоны водятся в Африке, Эта деревня расположена высоко в горах, Город лежит в долине Ганга, Паспорт куда-то делся*, но недопустимо: **Слоны водятся, *Эта деревня расположена, *Город лежит, *Паспорт делся*. Тем самым, синтаксис наречий *задолго* и *незадолго* может быть описан вполне системно, без постулирования уникальных структур типа (39). Специфика данного случая состоит в том, что он описывается с помощью сразу двух правил: (а) синтаксического переподчинения семантического актанта и (б) обязательного выражения этого же семантического актанта.

Критерий наречного управления уточняется вполне естественным образом: описание должно быть экономным, т. е. не требовать введения уникальных правил.

Маркированные случаи типа (27)–(32), когда в высказывании отсутствует выражение семантического актанта ‘временной ориентир’ наречия *задолго* или *незадолго*, описываются несколько иначе: не требуется обязательного выражения данного семантического актанта (действительно, этот семантический актант наречия *задолго* или *незадолго* может быть выражен в предтексте и тогда не описывается четким морфосинтаксическим правилом). Тем не менее, и в этом случае и наречие *задолго* или *незадолго*, и предложно-падежная группой *до* + РОД или *перед* + ТВОР подчиняются глаголу, т. е. и здесь имеет место обязательное синтаксическое переподчинение глаголу семантического актанта наречия.

Рассматриваемые наречия интересны еще в одном отношении. Коль скоро эти наречия обычно употребляются с последующей предложно-падежной группой, обозначающей их семантический актант ‘временной ориентир’, то, казалось бы, данные слова становятся подобны предлогам. Действительно, подобная ситуация имеет место с бывшими наречиями *между, среди, через* и т. п., которые в современном языке уже не употребляются без последующей падежной формы, а значит перешли в разряд предлогов.

Однако слова *задолго* и *незадолго*, даже если бы они не могли употребляться без постпозитивной предложно-падежной группы, все равно нельзя было бы трактовать как предлоги: данные наречия, в отличие от предлогов, можно опустить без нарушения грамматической правильности предложения. Кроме того, получилось бы, что данные производные предлоги противоречат определению наречного предлога. Действительно, в их составе не выделялось бы никакое наречие: поскольку по академической грамматике наречие не может требовать последующей предложно-падежной группы.

Литература

1. *Апресян 2010* — Апресян Ю. Д. Инструкция по составлению словарных статей Активного словаря (АС) русского языка // Проспект активного словаря русского языка. М.: Языки славянских культур, 2010. С. 156–152.
2. *Курилович 1962* — Курилович Е. Основные структуры языка: словосочетание и предложение // Курилович Е. Очерки по лингвистике. М.: Изд-во иностранной литературы, 1962. С. 48–56.

3. *РГ-70 — Русская грамматика / Гл. ред. Шведова Н. Ю. М.: Наука, 1970.*
4. *РГ 1980* — Русская грамматика: В 2 т. Т. I / Гл. ред. Шведова Н. Ю. М.: Наука, 1980.
5. *Урысон 2017* — Урысон Е. В. Предлог или наречие? Частеречный статус наречных предлогов // ВЯ, 2017. № 5. С. 36–55.
6. *Урысон 2014* — Урысон Е. В. О производных предлогах: наречные предлоги // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2014» (Бекасово, 2013 г.). Вып. 13 (20): в 2 т. М.: Изд-во РГГУ, 2014. Т. 1. С. 695–706.
7. *Холодович 1979* — Холодович А. А. К вопросу о доминанте предложения // Холодович А. А. Проблемы грамматической теории. Л., 1979. С. 293–298.

References

1. *Apresjan Ju. D.* (2010), Instruction on making lexical entries for Active Dictionary of Russian [Instrukcija po sostavleniju slovarnykh statej Aktivnogo slovaria (AS) russkogo jazyka]. Preliminaries for Active Dictionary of Russian [Prospekt aktivnogo slovaria russkogo jazyka]. Moscow, "Yazyki slavianskikh kul'tur", pp. 55–152.
2. *Kurilovich E.* (1962), Main language structures: word combination and clause [Osnovnyie structury jazyka: slovosochetanije i predlozhenije]. Essays in linguistics [Ocherki po lingvistike], Moscow: Foreign literature Publishing House, pp. 48–56.
3. *RG-70* (1970), Modern Russian Grammar of literary language [Grammatika sovremennogo russkogo literaturnogo jazyka], N. Ju. Shvedova (ed.), Moscow, Nauka.
4. *RG-80* (1980), Russian grammar [Russkaya grammatika]: In 2 vol. N. Yu. Shvedova (ed.). Moscow, Nauka.
5. *Uryson E. V.* (2017), An adverbial preposition or an adverb? [Predlog ili narechije? Chasterechnyj status narechnykh predlogov]. Problems of linguistics [Voprosy jazykoznanija], 5, pp. 36–55.
6. *Uryson E. V.* (2014), On derived prepositions: adverbial prepositions [O proizvodnykh predlogakh: narechnyje predlogi]. Computational Linguistics and Intellectual Technologies. Proceedings of the Annual International Conference "Dialogue" (2014), issue 13, v. 1, pp. 695–706.
7. *Kholodovich A. A.* (1979), To the question of the dominant of a sentence [K voprosu o dominante predlozhenija]. Problems of grammar theory [Problemy grammaticheskoj teorii], Leningrad, pp. 293–298.

ЧТО БУДЕТ, ТО (И) БУДЕТ: ОБ ОДНОМ КЛАССЕ ТАВТОЛОГИЧЕСКИХ КОНСТРУКЦИЙ В РУССКОМ ЯЗЫКЕ[1,2]

Вилинбахова Е. Л. (e.vilinbakhova@spbu.ru)
Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

В статье рассматриваются коррелятивные тавтологические конструкции вида *что будет, то (и) будет*, где придаточное предложение предшествует главному, а содержание обеих частей материально совпадает. При анализе материала из Национального корпуса русского языка и интернет-источников обнаруживается ряд нетривиальных особенностей, присущих данным конструкциям. Так, некоторые тавтологии в разных контекстах передают противоположные значения: *что было, то было* может интерпретироваться и как ‘то, что это действительно было, нельзя отрицать’ [Булыгина, Шмелев 1997], и как готовность забыть о прошлом в интересах будущего [Активный словарь русского языка]. Далее, частица *и* в главном предложении допустима в одних тавтологиях, но неприемлема в других. В работе предлагается объяснение указанным фактам путем выделения четырех возможных значений на основании двух оппозиций: (а) находится ли описываемая ситуация в фокусе внимания говорящего или выводится из него; (б) является ли прочтение конструкции генерическим или конкретно-референтным.

Ключевые слова: языковые тавтологии, коррелятивы, русский язык, микросинтаксис, семантика, прагматика

CHTO BUDET, TO (I) BUDET: ON ONE PATTERN OF TAUTOLOGIES IN RUSSIAN

Vilinbakhova E. L. (e.vilinbakhova@spbu.ru)
St. Petersburg State University, St. Petersburg, Russia

[1] Автор выражает признательность Кристине Крушинской за помощь в сборе материала и Сереже Саю за ценные наблюдения, а также анонимным рецензентам за их комментарии и предложения по доработке статьи.

[2] Работа поддержана грантом СПбГУ № 31.11.48.2017, грантом Президента Российской Федерации № МК-713.2017.6 и грантом Министерства экономики и конкурентоспособности Испании № FFI2015-64397-P (проект SPIRIM).

This paper contributes to the debate on the analysis of linguistic tautologies—structures that state an unquestionable truth by virtue of their logical form and therefore require a reinterpretation to be informative. While there is a great number of studies of nominal tautologies of the form '*X is X*', clausal tautologies, i. e. conditionals '*if P, P*', disjunctives '*either P or not P*', free relatives '*P, what P*', etc., are given less attention. This paper investigates one of such patterns, namely, correlative tautologies, where the subordinate clause precedes the main clause, that could be exemplified by the expression *chto budet to (i) budet* lit. 'what will be that (EMPH) will be'. The data taken from the Russian National Corpus and Internet as well as dictionary definitions show that tautologies of this kind exhibit various peculiar properties. First, some correlative tautologies can receive opposite interpretations in different contexts, i. e. *chto bylo, to bylo* lit.'what has been that has been' can mean both 'this fact cannot be denied' [Bylugina, Shmelev 1997] or 'the past should be forgotten for the sake of the future' [Active Dictionary of Russian]. Next, the particle *i*, which is commonly used in Russian correlatives, cf. [Mitrenina 2010], is acceptable for some tautologies but not licensed in others. I argue that for correlative tautologies the crucial ingredient is salience of the situation in question as presented by the speaker that, along with specific vs. generic readings available, results in four possible strategies of their interpretation.

Key words: tautologies, correlatives, Russian language, microsyntax, semantics, pragmatics

1. Вводные замечания

Языковые тавтологии вида *Жизнь есть жизнь* рассматриваются в лингвистических исследованиях уже более сорока лет (см. библиографию в [Rhodes 2009; Snider 2015]). Наиболее активно изучаются тавтологии вида *X cop X*, однако упоминаются и другие разновидности: условные *если P, то P*; взаимоисключающие *P или не P*; относительные *P, что P*, см. [Ward, Hirschberg 1991; Meibauer 2008].

В данной работе на материале Национального корпуса русского языка (далее — НКРЯ) и интернет-источников будут рассмотрены конструкции вида (1–3), где придаточное предложение предшествует главному, а содержание обеих частей материально совпадает.

(1) *Менеджеры стали безынициативны, один-двое что-то предлагали, остальные как кролики на удава смотрели: мол,* ***что будет, то и будет****.* [Наталья Филатова. Офис в кризисе. Как «трудовой планктон» переживает потрясения эпохи // «Русский репортер», № 25 (104), 02–09 июля 2009, 2009]

(2) *Кабанов расстарался и весьма красочно изобразил жаргон, детально — вооружение-снаряжение-амуницию спецназа и террористов.* ***Что знает, то знает****. Ни отнять, ни прибавить.* [Андрей Измайлов. Трюкач (2001)]

(3) *Я постоянна в своих симпатиях, тем более сложившихся близких и теплых отношениях и уж* ***кого люблю, того люблю***.[3]

В классификации относительных предложений А. А. Зализняка и Е. В. Падучевой данные примеры относятся к предложениям бессубстантивного класса архаического типа релятивизации (с инверсией подчинения), см. [Zaliznyak, Paducheva 1975]. Также в литературе представлены обозначения *местоименно-соотносительные сложноподчиненные предложения*, *примыкающие относительные предложения* и некоторые другие, см. [Belyaev 2014:116], а наиболее распространенным является термин *коррелятивные конструкции*, или *коррелятивы*, см., например, [Lyutikova 2008; Nikunlassi 2008; Mitrenina 2008, 2010, 2012; Belyaev 2012 и др. работы; Inkova 2013; Kholodilova 2014]. Таким образом, рассматриваемые здесь структуры будут обозначены как коррелятивные тавтологии (далее — КТ[4]).

Даже при беглом взгляде на примеры (1–3) бросаются в глаза различия в предаваемых оттенках значения. Так, в (1) КТ описывает равнодушие сотрудников и их нежелание участвовать в решении проблем компании, ср. с конструкцией *будь что будет*, которая выражает «граничащее с фатализмом безразличие к результату действия или бездействия говорящего» [Iomdin 2017:165]. В (2) КТ подчеркивает достоверность описываемого факта. Наконец, в (3) возможно прочтение КТ в качестве условной (что характерно для коррелятивов в целом, см., например, [Arsenijević 2009] и ссылки, приводимые в [Lypták 2009:26]) с указанием на полноту признака: ‘если я кого-то люблю, то люблю его в полную силу’; подтверждение любви к конкретным людям, упомянутым раньше, как в (2)[5]; выражение отношения субъекта к сообщаемому, как в (1): ‘люблю именно тех, кого люблю, и с этим ничего не поделаешь’ и т. д.

Также в (1) имеется частица *и*, которая, как отмечает О. В. Митренина, широко распространена в коррелятивных конструкциях в русском языке в позиции после коррелята; она «не является обязательной и может быть в большинстве случаев опущена, но тогда предложение звучит менее естественно»

---

[3] Примеры взяты из Интернета, если не указано иное. Из соображений краткости адреса не указаны, но, разумеется, в материалах они присутствуют. Орфография и пунктуация оригинала почти полностью сохранены.

[4] Следует оговорить, что в работе анализируются только относительные коррелятивы; за рамками остались примеры КТ со значением времени (i), количества (ii), и т. п.; думается, однако, что полученные выводы можно распространить и на указанные случаи.

(i) *Война, война… Это еще неизвестно, когда будет война.* ***Когда будет — тогда будет!*** *Это север!* [Константин Симонов. Беседы с Адмиралом Флота Советского Союза И. С. Исаковым (1962)]

(ii) *Радиоголовы выложили альбом с условием «****сколько заплатишь — столько заплатишь****»*

[5] См. (i), где объект любви называется:

(i) *Но вот возьмем мы, скажем, Людвига ван Бетховена! (****кого люблю — того люблю****!)*

[Mitrenina 2010:149, прим. 5] (перевод мой — ЕВ). Тем не менее, если мы подставим данную частицу в (2'), то как раз ее присутствие делает предложение менее естественным, а в (3') склоняет слушателя в пользу последней из указанных интерпретаций.

(2') *Кабанов <…> красочно изобразил жаргон, детально — вооружение-снаряжение-амуницию спецназа и террористов.* ***Что знает, то (??и) знает****. Ни отнять, ни прибавить.*

(3') *Я постоянна в своих симпатиях, <…> и уж* ***кого люблю, того и люблю****.*

В работе предпринимается попытка описать характеристики КТ в русском языке и дать возможные объяснения некоторым их свойствам. В разделе 2 дана история вопроса; в разделе 3 обсуждаются структурные и семантические свойства КТ; в заключении представлены итоги исследования.

2. История вопроса

Анализ КТ в литературе впервые[6] встречается в работах А. Вежбицкой, см. [Wierzbicka 1987, 1991]. Она рассматривает тавтологии с заполненными переменными — аналоги выражений *что было, то было* и *что будет, то будет* в ряде языков, и продуктивную синтаксическую модель в польском языке *Ale co S to S* 'Но что *S*, то *S*', отмечая, что в первом случае описываемая КТ ситуация оценивается как отрицательная, а во втором как положительная [Wierzbicka 1991:431–437].

В [Sonnenhauser 2017] рассматривается немецкая КТ *Wer kann, der kann* букв. 'Кто может, тот может', у которой выделяется две дискурсивные функции: усиление высказываемой мысли (англ. *reinforcement*), или ее обоснование (англ. *justification*).

На русском материале Т. В. Булыгина и А. Д. Шмелев отмечают, что польский оборот *Co było to było* соответствует русскому *что было, то сплыло; что было, то прошло,* а не формальному аналогу *что было, то было* со значением 'то, что это действительно было, нельзя отрицать', ср. с оборотами *что правда, то правда* и *что есть, то есть* [Bulygina, Shmelev 1997: 445]; см. также [Paducheva 1991] и [Krushinskaya 2017].

Некоторые КТ в русском языке фиксируются в словарях, см.:

---

[6] Из соображений краткости история вопроса приводится в сжатом виде, однако следует отметить, что КТ мельком упоминаются в ряде публикаций, посвященных т. н. «сравнительным тавтологиям» (вроде *The fact that John is as tall as he is disturbs me* букв. 'Тот факт, что Джон такой высокий, какой он есть, меня беспокоит') и их возможным прочтениям de re и de dicto, в 1970-х гг в журнале «Лингвистические исследования» (Linguistic Inquiry), см. библиографию в [Horn 1981], а также в [Ward, Hirshberg 1991, Rhodes 2009].

Что было, то было [Обычно в прямой речи] Говорящий заявляет, что готов забыть о том плохом, что было в прошлом, в интересах будущего[7] [ADR: 400].

Что есть, то есть *Разг. Экспрес.* Согласен; действительно так [Fedorov 2008].

Что будет, то будет [Обычно в прямой речи] Говорящий заявляет, что готов принять любой исход имеющих место событий [ADR: 400]; *разг.* О действии, предпринимаемом наудачу [Fedorov 2008].

Также даны КТ ***что выросло, то выросло*** [Serov 2003], ***что правда, то правда*** [Ozhegov, Shvedova 1992], ***что верно, то верно*** и ***что да, то да*** в словаре синонимов [Trishin 2013] в ряду с выражениями *истинная правда, так оно и есть, подлинно* и пр., ***чего нельзя того нельзя***, ***да уж я что знаю, то знаю*** в «Пословицах русского народа» [Dal’ 1879/2000].

Таким образом, в литературе чаще всего предлагается описание КТ с заполненными переменными, которые имеют статус устойчивых и фиксируются в словарях. В данной работе рассматриваются в первую очередь продуктивные модели КТ.

3. Общие свойства коррелятивных тавтологий в русском языке

3.1. Структурные свойства КТ

В структурном отношении встреченные в материалах КТ демонстрируют большее разнообразие, чем КТ, описанные в литературе и словарях: так в рассмотренных выше примерах чаще всего представлена релятивизация подлежащего и прямого дополнения, хотя в КТ релятивизируются и другие синтаксические позиции, см. (4–5).

(4) *И не важно, что только троим из нас достанутся премии*. ***Кому дадут, тому дадут***. *Главное, что все достойны, работы всех замечены.*

(5) ***Без чего не могу, без того не могу***… — *довольным голосом проговорил Дэго, забирая книжку.*

Также встретились КТ, где релятивизируются более одной позиции, что является одним из характерных свойств коррелятивов в целом [Lipták 2009: 5], см. (6). Поскольку мы имеем дело с тавтологиями, и в простых, и во множественных КТ происходит релятивизация одних и тех же позиций — условие, необязательное для прочих «нетавтологических» коррелятивов, см. (7).

[7] Как видно, данная трактовка отличается от указанной в [Падучева 1991] и [Булыгина, Шмелев 1997] и соответствует польскому аналогу.

(6) *Давайте* ***кто что сделает, тот то и сделает****. А голосовать бум [будем] за конкретных кукол.*

(7) *Ну / кому что надо / тот то и приватизировал*. [Беседа в Новосибирске (2000.08.15)] (пример из [Лютикова 2008:23, № 168])

Отметим, что в КТ полное совпадение главного и придаточного предложения встречаются далеко не всегда. Возможна ситуация, когда в первой «копии» представлено полное предложение, а во второй — только предикативная вершина[8], см. (8–9). Таким образом, полное совпадение — это частный случай, когда подчиненное предложение совсем короткое.

(8) *Но* ***что Петров накаркал, то накаркал****: ни единого всхода не дал посев, словно это Молочкову приснилось, что он сажал в мае месяце лук-порей.* [Вячеслав Пьецух. Летом в деревне // «Новый Мир, 2000]

(9) ***Что было в Уфе, то было****, уже ничего не изменить. Нужно начинать всё заново.*

В целом, в структурном аспекте КТ укладываются в рамки существующих описаний коррелятивов в русском языке с учетом отдельных особенностей (например, невозможность релятивизации несовпадающих синтаксических позиций в главном и придаточном предложениях), следующих из их тавтологической природы.

3.2. Семантические свойства КТ

Обращаясь к семантике КТ, следует отметить, что коррелятивные конструкции относятся к особому семантическому классу относительных предложений, который выделяется наряду с классами рестриктивных (где придаточное предложение ограничивает экстенсионал именной группы в главном) и аппозитивных (где этого ограничения не происходит) предложений и обозначается как «максимализирующий» (термин введен в [Grosu, Landman 1998]), т. е. «ситуация в зависимой клаузе с необходимостью оказывается верна для всего множества участников, описываемых относительной конструкцией» [Kholodilova 2014:§3.2.1]. Как пишет О. В. Митренина, это справедливо и для русских коррелятивов, которые могут отсылать либо к конкретному объекту, либо ко всему множеству объектов в целом [Mitrenina 2012:64].

Для встреченных ранее КТ, где речь идет о конкретном референте, см. уже приводимый выше пример (10), можно предложить следующее толкование: ‘Говорящий утверждает, что в отношении участника X имеет место ситуация P[9]; ситуация P реализуется в полной мере’.

(10) *— Но вот возьмем мы, скажем, Людвига ван Бетховена!* ***(кого люблю — того люблю****!)*

[8] Автор благодарен за это наблюдение С. Саю.

[9] Термины *участник* и *ситуация* используются в трактовке [Paducheva 2004:52].

Следует отметить, что давать характеристику конкретному референту могут и тавтологии, устанавливающие тождество, см. (11).

(11) *И, учитывая инф[ормацию] о религиозности Баха, это всё для него были не пустые слова и он старался (несмотря на то, что выдавал на гора по кантате в неделю, но тут уж* ***Гений есть Гений****), чтобы кантаты выполняли это «целеполагание»* ‘гений — это Бах’

Однако, как отмечает Е. В. Падучева, в данном случае категоризация не есть ассерция:

> Сложность смысла высказывания *X есть X* обусловлена тем, что содержательно главный компонент ее значения — x относится к категории X — присутствует в ее толковании в статусе За кадром; участник x по поводу которого делается высказывание в поверхностной структуре не отражен [Падучева 2004: 105].

Таким образом, если анализировать КТ по аналогии с тавтологиями *X есть X*, то в подобных примерах следует постулировать генерическое прочтение, когда происходящая с референтом ситуация выступает как конкретная манифестация утверждаемой закономерности.

В пользу такой трактовки говорит пример (12)[10] из романа Марининой, где в диалоге следователя и подозреваемого, который рассказывает на допросе, как он похитил некую жертву, в какой-то момент возникают следующие реплики:

(12) *— После субботы я этого кренделя в глаза не видел.*
— Ну да, — кивнул Антон, — а в воскресенье убил его. За что, Гриша? Что он тебе сделал?
—Уби-ил? — недоверчиво протянул Дубинюк. — Э, нет, начальник, так не пойдет. ***Что мое — то мое****, упираться не стану, вон командир велел признаваться — его слово закон. А чужого мне не шей.*
[Александра Маринина. Последний рассвет (2013)]

Далее КТ *что мое — то мое* повторяется еще несколько раз. В данном случае, КТ одновременно отсылает к конкретному факту похищения, в котором и признается подозреваемый, и используется героем как общая характеристика ‘я признаю все свои проступки’, что увеличивает доверие к отрицанию более сильного утверждения — обвинения в убийстве.

И все же думается, что в случае с КТ данные интерпретации следует разделять, т. к. они отличаются наличием (для генерического прочтения) смыслового компонента ‘возможно’, т. е. гипотетической возможности нереализации описываемой ситуации, общего с условными предложениями, см. толкование союза *если*, например, в [Санников 2008:414 и след.; Урысон 2011:15 и след.]. Так, есть вероятность, что в (3) героиня одинока, и ситуация (настоящей) любви не имеет места в действительности из-за отсутствия любимых людей,

[10] Автор благодарен за привлечение данного примера в качестве аргументации С. Саю.

а автодескрипция *что мое — то мое*, произнесенная до ситуации признания похищения, остается без конкретных подтверждений. Прочие компоненты толкования сохраняются.

Тем не менее, для некоторых примеров данного анализа оказывается недостаточно: он не учитывает отношения говорящего к сообщаемому, которое передается в (1) и (13–14).

(13) *Будем атаковать. Только чур без обид.* ***Кого выберет, того выберет.***
[Олег Дивов. Молодые и сильные выживут (1998)]

(14) *Два дня продолжалась лихорадочная переработка добытого.* ***Что успели, то и успели.*** [Владимир Солоухин. Третья охота (1967)]

Из компонентов значения КТ, представленных в литературе и словарях, к данным примерам подходит: ‘готовность принять любой исход имеющих место событий’ из толкования *что будет, то будет* в [ADR:400], а также часть толкования *будь что будет,* не относящейся к КТ, ‘безразличие к результату’, см. [Iomdin 2017:165].

Кажется, эти компоненты и указанное выше толкование КТ строятся по разным принципам (описание отношения говорящего vs. описание ситуации), т. е. не являются взаимоисключающими, но в ряде случаев они оказываются прямо противопоставлены друг другу. Например, как было упомянуто в разделе 1, в КТ с выражением отношения используется или, по крайней мере, может быть добавлена частица *и*, а в КТ, описывающих ситуацию, она чаще всего не допускается, см. (2’) и (12’).

(12’) *Э, нет, начальник, так не пойдет.* ***Что мое — то (??и) мое****, упираться не стану,*

Исходя из существующих описаний частицы *и*, см. [Shimchuk, Shchur 1999, Uryson 2011], можно выделить следующие наблюдения, которые, как представляется, имеют отношение к рассматриваемым примерам: указание частицы *и* на тождество объектов или ситуаций [Shimchuk, Shchur 1999:69; Uryson 2011:271 и след.] и несовместимость с выражениями *совершенно, полностью, начисто* [Uryson 2011:270].

Первое свойство, указание на тождество, казалось бы, полностью соответствует тавтологической природе рассматриваемых конструкций. Тем не менее, при более внимательном анализе оказывается, что КТ, описывающие ситуацию, являются т. н. псевдотавтологиями, поскольку второй повторяющийся элемент привносит новую информацию для слушателя, указывая на достоверность события и полноту его реализации, см. возможность опровержения (4), повторенного как (15a), в (15b).

(15) a. — ***Без чего не могу, без того не могу****… — довольным голосом проговорил Дэго, забирая книжку.*

b. *Это неправда, — возразил Дэви* [собеседник героя в цитируемом рассказе — ЕВ] — *ты без нее прекрасно обходился все это время / ты бы и не вспомнил о ней, если бы не я* и т. д.

Напротив, в КТ, описывающих отношение говорящего, тождество между элементами является полным, поэтому отрицанию подвергается импликатура, а не пропозициональное содержание, см. (16), где автор возражает против беспечного отношения адресата к последствиям своих действий.

(16) *В след[ующий] раз думай что из ситуации может получится, а не что будет то будет.*

Второе свойство — неестественность употребления частицы *и* во фразах типа (17), т. е., несовместимость с компонентом 'ситуация реализуется в полной мере', указывает, что данный компонент как будто не должен использоваться в КТ, выражающих, скажем, безразличное отношение говорящего[11].

(17) ??*Ты меня извини, я про это совершенно (полностью, начисто) и забыл*
[Uryson 2011:270, № 18]

Заметим, что данное противопоставление присутствует не только в КТ, но и в двух омонимичных конструкциях с тождественными словоформами, обозначенными в [Kopotev, Faynveyts 2007] как *X так X!* 'полноты признака' и *X так X* 'неконтролируемого выбора', см. возможные ответы (18b) и (18c).

(18) a. *Пойдем сегодня гулять?*
b. *Конечно! Поедем за город! Гулять так гулять!*
c. *Мне все равно… Гулять так гулять…*

Можно предположить, что такое противопоставление связано с тем, считает ли говорящий нужным удерживать описываемую ситуацию в фокусе внимания. В примерах КТ, выступающих как подтверждение истинности ситуации и указание на полноту ее реализации, ситуация находится в фокусе внимания говорящего[12]. В прочих примерах говорящий указывает, что выводит ситуацию из фокуса своего внимания и/или призывает сделать это адресата, поэтому такие примеры нередко выступают маркером закрытия темы. Если же ситуация

---

[11] По наблюдению анонимного рецензента, на допустимость употребления в КТ частицы *и* также может влиять время глагола, например «что случится, то и случится» лучше, чем «что случилось, то (?и) случилось». Оставляя этот вопрос для дальнейшего исследования, отмечу, что в материалах подобные примеры в прошедшем времени присутствуют, см. (i).

(i) *«Значит, Господь такой жертвы потребовал». Что случилось, то и случилось.*
[Юрий Буйда. У кошки девять смертей (2000) // «Новый Мир», 2005]

[12] Анонимный рецензент поднял вопрос о способах проверки того факта, что ситуация находится в фокусе внимания. Кажется, одним из таких тестов может быть возможность интонационной и графической выделенности второго повторяющегося элемента, и ее меньшая приемлемость для КТ, исключающих ситуацию из фокуса внимания, см. (10) и (14), повторенные здесь как (i) и (ii):

(i) *— Но вот возьмем мы, скажем, Людвига ван Бетховена!* (***кого люблю — того*** OK***ЛЮБЛЮ***!)

(ii) *Два дня продолжалась лихорадочная переработка добытого.* ***Что успели, то и*** ??***УСПЕЛИ***.

исключается из фокуса внимания, то полнота её реализации уже не имеет значения, поэтому данный смысловой компонент не встречается в соответствующих примерах.

Следует указать, что основанием такого исключения чаще всего является неконтролируемость описываемой ситуации (этот компонент характерен для тавтологий в целом), а сопутствующим передаваемым отношением оказывается безразличие к ее последствиям, готовность их принять и т. д.

Поскольку КТ со значением исключения ситуации из фокуса внимания также могут иметь конкретно-референтное и генерическое прочтения, различающиеся компонентом ‘возможно’, можно выделить четыре типа КТ, проиллюстрированные примерами (19–22), и сформулировать следующие толкования:

1) $КТ_1$ ‘Говорящий утверждает, что в отношении участника X имеет место ситуация P;
ситуация P реализуется в полной мере;
ситуация P находится в фокусе внимания говорящего’, см. (19), а также (2, 5, 8, 10, 12)

(19) *Третье место. Жиркову за дриблинг, позволивший вывернуться из тисков двух соперников. Правда, проделал сие глубоко в середине поля. Но* ***что умеет, то умеет*** *— крепко освоенный навык не вязнет даже в рыхлом весеннее газоне.* [Юрий Цыбанев. Когда в ногах правды нет. Юрий Цыбанев попытался найти красоту в матче «Локо» — «Анжи» // Советский спорт, 2012.03.19]

2) $КТ_2$ ‘Говорящий утверждает, что возможна ситуация P, возможна ситуация не-P;
говорящий представляет, что имеет место ситуация P;
говорящий утверждает, что в рамках этой гипотезы (а) ситуация P с необходимостью верна для всего множества участников X-ов; (б) ситуация P реализуется в полной мере;
ситуация P находится в фокусе внимания говорящего’, см. (20), а также (3, 12).

(20) *Сара* ***—*** *ровесница дочки, характера довольно крутого. Но* ***с кем дружит, с тем дружит****, и тут кремень.*

3) $КТ_3$ ‘Говорящий утверждает, что в отношении участника X имеет место ситуация P;
говорящий считает правильным вывести ситуацию P из фокуса внимания вследствие ее неконтролируемости[13], см. (21), а также (9, 14).

[13] Отмечу, что для $КТ_3$ в будущем времени набор контекстов невелик, вроде гипотетической ситуации, когда будущие родители узнают при обследовании, что родится ребенок «не того» пола, и произносят (i).
(i) *Жаль, конечно, но уж кто родится, тот и родится.*

(21) *Вообще было ошибкой брать ее с собой, но* ***что сделано, то сделано***. [Михаил Гиголашвили. Чертово колесо (2007)]

4) $КТ_4$ ‘Говорящий утверждает, что возможна ситуация Р, возможна ситуация не-Р;
говорящий представляет, что имеет место ситуация Р;
говорящий утверждает, что в рамках этой гипотезы ситуация Р с необходимостью верна для всего множества участников Х-ов;
говорящий считает правильным вывести ситуацию Р из фокуса внимания вследствие ее неконтролируемости’, см. (22), а также (1, 3’, 4, 6, 13).

(22) *Надо с ребенком ходить на выставки.* ***Что увидел, то увидел****. Это как слушать музыку. Надо слушать музыку, чтобы ее понимать. Много слушать музыки.* [Ольга Андреева, Ирина Антонова. Жду зеленых листочков // «Русский репортер», 2014]

Следует также признать, что встретились КТ, которые не вполне укладываются в данные четыре типа[14]: в первую речь идет о примерах вроде (23), где интерпретация ‘одобряет / молчит в полной мере’ или ‘следует вывести ситуацию одобрения / молчания из фокуса внимания коммуникантов’ не вполне естественна. Представляется, что в подобных случаях не всегда можно говорить о нахождении в фокусе внимания говорящего (либо выведения из него) каждой из ситуаций по отдельности, поскольку в ряде случаев подчеркиваются прежде всего различия между ними[15], хотя, разумеется, указанные значения могут и сохраняться.

(23) *Некоторые могут подумать, что замалчивание не есть критика, но это неправильно,* ***ибо кто одобряет, тот одобряет****, а* ***кто молчит, тот молчит***. [Аркадий Львов. Двор (1981)]

Количественное распределение продуктивных относительных КТ в основном корпусе НКРЯ (1945–1917 гг. создания текстов) дано в **Таблице 1**. Из рассмотрения были исключены лексикализованные КТ, указанные в разделе 2[16], а также примеры, где используются разные языковые выражения[17]. Поскольку

---

[14] Автор благодарен анонимному рецензенту за предложение внести данный комментарий.

[15] То же справедливо для конструкций *X есть X* и *X это X*, когда в составе множественных сопоставительных тавтологий происходит нейтрализация их семантических различий, см. [Vilinbakhova 2016], ср. также перифразу (23) в (i):
(i) *Одобрение — это одобрение, а молчание — это молчание.*

[16] Можно заметить, что значения лексикализованных КТ, отмеченные в литературе и словарях, укладываются в приведенные выше толкования: *что было, то было* в трактовке из [Булыгина, Шмелев 1997], а также *что есть, то есть*, *что правда, то правда*, и пр. относятся к $КТ_1$, но могут использоваться и как $КТ_2$; *что будет, то будет* и *что было, то было* по [АСРЯ 2014] могут употребляться как $КТ_3$ и $КТ_4$. О (не) соответствии значений лексикализованных КТ в текстах основного корпуса НКРЯ (1945–1917 гг. создания) словарным толкованиям см. [Krushinskaya 2017].

[17] (i) *Поэтому Шершеневич любил повторять крылатую фразу Мережковского: «Что по́шло, то и пошло́».* [Анатолий Мариенгоф. Мой век, мои друзья и подруги (1956–1960)]

некоторые КТ одновременно отсылают к конкретному факту и выступают в качестве общей характеристики, как в примере (12) из романа А. Марининой, количество КТ в разных прочтениях превышает количество собранных примеров (один пример может относиться к $КТ_1$ и к $КТ_2$).

В целом, можно отметить, во-первых, что КТ, описывающие ситуацию в фокусе внимания говорящего, и КТ, где ситуация выводится из него, встречаются в рассмотренных текстах практически в равной степени. Далее, большая часть примеров имеет конкретно-референтное прочтение, т. е. описывает отдельно взятую ситуацию. Наконец, представленность в корпусе сопоставительных КТ позволяет расширить категорию сопоставительных тавтологий в целом, и включить в качестве ее составляющих не только тавтологии, устанавливающие тождество, как это принято в предшествующих работах, но и пропозициональные тавтологии.

Таблица 1

Интерпретация	Продуктивные КТ
$КТ_1$ (ситуация в фокусе внимания, конкретно-референтное прочтение)	51
$КТ_2$ (ситуация в фокусе внимания, генерическое прочтение)	22
$КТ_3$ (ситуация выводится из фокуса внимания, конкретно-референтное прочтение)	57
$КТ_4$ (ситуация выводится из фокуса внимания, генерическое прочтение)	18
ДРУГОЕ (в частности, «нейтрализация», возможная в рамках сопоставительных КТ)	13
Σ	157

4. Заключение

В работе на материале НКРЯ и интернет-источников были рассмотрены структурные и семантические свойства тавтологических конструкций вида *что будет, то (и) будет*, обозначенных как коррелятивные тавтологии. Было выделено четыре типа конструкций на основании двух оппозиций: (а) находится ли описываемая ситуация в фокусе внимания говорящего или выводится из него; (б) является ли прочтение конструкции генерическим или конкретно-референтным.

Предложенное описание может быть полезным для задач автоматической обработки текста, в том числе RTE (Recognizing Textual Entailment ‘логический вывод по фрагменту текста’) и машинного перевода, в следующих аспектах[18].

[18] Автор благодарен анонимному рецензенту, который затронул данный вопрос, за идею добавить соответствующий комментарий.

Во-первых, полученные данные важны для анализа коммуникативной структуры соответствующих высказываний (о коммуникативной структуре см., например, [Melchuk 2001; Yanko 2001] и др.). В частности, можно предположить, что один из типов, $КТ_1$, является реализацией т. н. фокуса *verum*, который фиксирует внимание адресата на достоверности описываемой пропозицией ситуации и исключении противоположной альтернативы ее неосуществления, см. библиографию в [Leonetti, Escandell-Vidal 2009]. Отмечу, что в английском и испанском языках аналогичная синтаксическая конструкция для выражения данного значения невозможна. Далее, описание структурных особенностей, и в том числе ограничений для ряда КТ, могут помочь при выведении правильных инференций для данных конструкций (о роли инференций, основанных на вероятных ожиданиях, для автоматической обработки текста на примере косвенных речевых актов, см. [Boguslavski et al. 2016]). Наконец, представленные данные могут внести вклад в моделирование общих процессов интерпретации семантически аномальных в буквальном смысле языковых выражений, которые, тем не менее, активно используются в коммуникации.

Литература

1. *Active dictionary of Russian* [Aktivnyj slovar' russkogo jazyka], Ju. D. Apresyan (ed.), Moscow, 2014, vol. 1–2.
2. *Arsenijević B.* (2009), {Relative {conditional {correlative clauses}}}, in Lipták A. (ed.) Correlatives cross-linguistically, John Benjamins, Amsterdam; Philadelphia, pp. 131–156.
3. *Belyaev O. I.* (2012), Correlative construction and relative sentences with an inner head in the Besermyan dialect of the Udmurt language [Korrelyativnaya konstrukciya i otnositel'nye predlozheniya s vnutrennej vershinoj v besermyanskom dialekte udmurtskogo yazyka], in Kuznecova A. I., Toldova S. Ju., Saj S. S., Kalinina E. Ju. (eds.), Finno-Ugric languages: fragments of the grammatical description. Formal and functional approaches [Finno-ugorskie yazyki: fragmenty grammaticheskogo opisaniya. Formal'nyj i funkcional'nyj podhody], Manuscripts of Ancient Rus' [Rukopisnye pamyatniki Drevnej Rusi], Moscow, pp. 647–679.
4. *Belyaev O. I.* (2014), Book review of: Inkova, O., Hadermann, P. (eds.), (2013), Correlation: Syntactic and Semantic Aspects [La corrélation: Aspects syntaxiques et sémantiques], Librairie Droz S. A., Genève, in Topics in the Study of Language [Voprosy yazykoznaniya], Vol. 6, pp. 116–125.
5. *Boguslavski I., Dikonov V., Frolova T., Iomdin L., Lazurski A., Rygaev I., Timoshenko S.* (2016), Plausible Expectations-Based Inference for Semantic Analysis, in Proceedings of the International Conference on Artificial Intelligence (ICAI), CSREA Press USA, Las Vegas Nevada, USA, pp. 477–483.
6. *Bulygina T. V., Shmelev A. D.,* (1997), Linguistic conceptualization of the world (on the material of the Russian grammar) [Yazykovaya kontseptualizatsiya mira (na materiale russkoy grammatiki)], Languages of Slavic Culture [Jazyki slavyanskoy kul'tury], Moscow.

7. *Dal' V.I.* (1879 / 2000), Poslovitsy russkogo naroda [Proverbs of the Russian people], Olma-Press, Moscow.
8. *Grosu A., Landman F.* (1998), Strange Relatives of the Third Kind, Natural Language Semantics, Vol. 6 (2), pp. 125–170.
9. *Fedorov A. I.* (2008), Phraseological Dictionary of the Russian Literary Language [Frazeologicheskii slovar' russkogo literaturnogo yazyka], Astrel': ACT, Moscow.
10. *Horn, L. R.* (1981), A pragmatic approach to certain ambiguities, Linguistics and Philosophy, Vol. 4, pp. 321–358.
11. *Inkova O.* (2013), Domains of correlation in Russian [Domaines de corrélation en russe], in Inkova O., Hadermann P. (eds.), The Correlation: Semantical and Syntactical Aspects [La corrélation: Aspects syntaxiques et sémantiques], Librairie Droz, Genève.
12. *Iomdin L. L.* (2017), What to do about constructions like what to do [Kak nam byt' s konstruktsiyami tipa kak byt'], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2017"], Moscow, Vol. 2, pp. 161–175.
13. *Kholodilova M. A.* (2014), Relative Clauses [Otnositel'nye predlozhenia], Materials for corpus-based grammar of Russian [Materialy dlya proyekta korpusnogo opisaniya russkoy grammatiki], available at http://rusgram.ru/Относительные_придаточные
14. *Kopotev M. V., Faynveyts A. V.* (2007), Izuchat' tak izuchat': synchrony and diachrony [Izuchat' tak izuchat': sinkhroniya i diakhroniya], Scientific and technical information. Series 2, Informational processes and systems [Nauchno-tekhnicheskaya informatsiya, Seriya 2, Informatsionnye protsessy i sistemy], Vol. 9, pp. 29–37.
15. *Krushinskaya K. S.* (2017), Tautologies in complex sentences on the material of the Russian language [Tavtologii v slozhnopodchinennyh predlozheniyah na materiale russkogo yazyka], diploma paper, Saint Petersburg State University, Saint Petersburg.
16. *Leonetti M., Escandell-Vidal V.* (2009), Fronting and verum focus in Spanish, in Dufter, A., Jacob, D. (eds.), Focus and background in Romance languages, Amsterdam: John Benjamins, pp. 155–204.
17. *Lipták A.* (2009), The Landscape of Correlatives, in Lipták A. (ed.), Correlatives cross-linguistically, John Benjamins, Amsterdam; Philadelphia, pp. 1–48.
18. *Lyutikova E. A.* (2008), Riddles of Russian relative sentences [Zagadki russkih otnositel'nyh predlozhenij], paper presented at Syntactic structures-2 [Sintaksicheskie struktury-2], Moscow.
19. *Meibauer J.* (2008), Tautology as Presumptive Meaning, Pragmatics and Cognition, Vol. 16, pp. 439–470.
20. *Melchuk I. A.* (2001), Communicative Organization in Natural Language: The semantic-communicative structure of sentences. Benjamins Academic Publishers, Amsterdam.
21. *Mitrenina O. V.* (2008), Syntax of correlative constructions in Russian: a generative approach [Sintaksis korrelyativnikh konstruktsiy russkogo yazika s pozitsii generativnoy grammatiki], Computational Linguistics and Intellectual

Technologies: Proceedings of the International Conference "Dialogue 2008" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2017"], Moscow, pp. 356–361.
22. *Mitrenina O. V.* (2010), Correlatives: Evidence from Russian, Formal Studies in Slavic Linguistics: Proceedings of Formal Description of Slavic Languages, 7.5, pp. 135–152.
23. *Mitrenina O. V.* (2012), The Syntax of Pseudo-Correlative Constructions with the Pronoun Kotoryj ('Which') in Middle Russian, Slověne, Vol. 1, pp. 61–73. Available at: www://slovene.ru/2012_1_Mitrenina.pdf
24. *Nikunlassi A.* (2008), Adnominal relative constructions in contemporary Russian [Primestoimenno-otnositel'nye konstrukcii v sovremennom russkom jazyke]. PhD thesis, University of Helsinki., Helsinki University Press, Helsinki.
25. *Ozhegov S. I., Shvedova N. Yu.* (1992), Dictionary of the Russian Language [Tolkovy slovar' russkogo jazyka], Az, Moscow.
26. *Paducheva E. V.* (1997), The phenomenon of Anna Wierzbicka, in: Wierzbicka A., Semantics, Culture, and Cognition [Yazyk. Kul'tura. Poznanie], Russian dictionaries [Russkie slovari], Moscow, pp. 5–32.
27. *Paducheva E. V.* (2004), Dynamic models in lexical *semantics* [Dinamicheskie modeli v semantike leksiki], Languages of Slavic culture [Yazyki slavyanskoy kul'tury], Moscow.
28. *Rhodes R.* (2009), A Cross-linguistic Comparison of Tautological Constructions with Special Focus on English, available at: www.linguistics.berkeley.edu/~russellrhodes/pdfs/taut_qp.pdf
29. *Russian National Corpus* [Natsional'nyy korpus russkogo yazyka], available at: www.ruscorpora.ru
30. *Sannikov V. Z.* (2008), Russian syntax in a semantic-pragmatic context [Russkij sintaksis v semantiko-pragmaticheskom prostranstve], Languages of Slavic cultures [Jazyki slavyanskih kul'tur], Moscow.
31. *Shimchuk E., Shchur M.* (1999) Dictionary of Russian particles [Slovar' russkih chastic], Berliner Slavistic works [Berliner slavistishe Arbeiten], Vol. 9, Frankfurt am Main.
32. *Serov V. V.* (2003), The Encyclopedic dictionary of proverbial words and expressions [Enciklopedicheskiy slovar krilatih slov i virazheniy], Lokid-Press, Moscow.
33. *Snider T.* (2015), Using tautologies and contradictions, Proceedings of the 19th Sinn und Bedeutung, Vol. 19, pp. 590–607.
34. *Sonnenhauser, B.* (2017), Tautologies at the interfaces: Wer kann, der kann, Journal of Pragmatics, Vol. 117, pp. 16–28.
35. *Trishin V. N.* (2013), Dictionary Catalog of Russian Language Synonyms [Bolshoy russky slovar'-spravochnik sinonimov], available at trishin.net
36. *Uryson E. V.* (2011), Description of the Semantics of Conjunctions: Language Data on Conscience Activity [Opyt Opisaniia Semantiki Soiuzov: Dannye Iazyka o Deiatel'nosti Soznaniia], Yazyki Slavyanskikh Kultur, Moskva.
37. *Vilinbakhova E. L.* (2016), Coordinated tautologies in Russian [Sopostavitelnyie tavtologii s russkom yazyke], Topics in the study of language [Voprosy Jazykoznanija], Vol. 2, pp. 61–74.

38. *Ward, G. L., Hirschberg J.* (1991), A Pragmatic Analysis of Tautological Utterances, Journal of Pragmatics, Vol. 15 (6), pp. 507–520.
39. *Wierzbicka A.* (1987), Boys Will Be Boys: 'Radical Semantics' vs. 'Radical Pragmatics', Language, Vol. 63 (1), pp. 95–114.
40. *Wierzbicka A.* (1991), Cross-Cultural Pragmatics: The Semantics of Human Interaction, Mouton de Gruyter, Berlin; New York.
41. *Yanko T. E.* (2001), Communicative strategies of Russian speech [Kommunikativnye strategii russkoj rechi], Yazyki Slavyanskikh Kultur, Moskva.
42. *Zaliznyak A. A. Paducheva E. V.* (1975), Towards the typology of relative sentences [K tipologii otnositel'nogo predlozheniya], Semiotics and Informatics [Semiotica i informatica], Vol. 6., pp. 51–101.

РЕЧЕВЫЕ АКТЫ В СТРУКТУРЕ СВЯЗНОГО ДИСКУРСА: ПОКАЗАТЕЛИ НЕЗАВЕРШЕННОСТИ ПО ДАННЫМ КОРПУСОВ ЗВУЧАЩЕЙ РЕЧИ[1]

Янко Т. Е. (tanya_yanko@list.ru)
Институт языкознания РАН, Москва, Россия

IMPERATIVES, VOCATIVES, AND QUESTIONS IN COHERENT DISCOURSE: THE PROSODIC MARKERS OF INCOMPLETENESS IN THE RUSSIAN SPOKEN SPEECH CORPORA

Yanko T. E. (tanya_yanko@list.ru)
Institute of Linguistics, Moscow, Russia

One of the means of designating the coherence in the spoken discourse is demonstrating that the current utterance of the discourse is not terminal. Every step of narrative consisting of the chain of statements can be marked as non-final. The prosodic cues for incompleteness applied to the speech act of a statement have been studied in details in linguistic literature. In this paper, the discourse incompleteness is analyzed as composed not only with statements but with questions, imperatives, and vocatives as well. The results of the investigation are as follows. The *wh*-questions, imperatives, and vocatives can be freely composed with the meaning of discourse continuity, and they have specific prosodic cues for marking this combination of meanings. Whereas the *yes-no*-questions do not accept the prosodic incompleteness marking. The prosodic patterns of incompleteness and the accent placement in questions, vocatives, and imperatives are exemplified here by the dialogues taken from the Multimodal corpus of the Russian National corpus, the Prosodically Annotated Corpus of Spoken Russian (spokencorpora.ru), and the minor working collection of the Russian speech recordings specifically set up for this investigation. The software program Praat was used in the process of analyzing the sounding data.

Keywords: pitch accents, prosody, discourse, incompleteness, accent-placement, Praat, spoken language, Russian, the Russian National corpus, statement, question, vocative, imperative

[1] Исследование выполнено в ФГБУН Институт языкознания РАН при поддержке Российского научного фонда (РНФ), проект №14-28-00130.

Один из способов поддержания связности дискурса — это указание на то, что продолжение повествования следует. Каждый шаг повествования, состоящего из ряда сообщений, может быть оформлен как неконечный. Просодические средства поддержания связности устного русского нарратива изучены достаточно детально ([Bryzgunova 2003: 895], [Yanko 2008: 128–170], [Kibrik, Podlesskaja 2009: 96–176], [Podlesskaya V. 2016], [Yanko 2017]). Основное просодическое средство поддержания связности — различные типы повышения частоты основного тона. Между тем повышения частоты создают эффект незавершенности не сами по себе, а только будучи параметром определенной словоформы — акцентоносителя незавершенности. Выбор словоформ — носителей значения незавершенности и других коммуникативных значений (темы, ремы, вопросительности и других) — не случаен; он подчинен особым принципам, см. [Yanko 2008: 49]. Так, в рассказе пострадавшей от взрыва бытового газа подъемы частоты основного тона на ударных слогах словоформ *комнату* и *стекло* говорят о том, что текущий шаг повествования не последний:

(1) *Потом я подошла в другую* ***ко/мнату****, вот у меня выбито* ***стекло/****, но запаха газа не* ***ощуща\ю***[2] (звучащая запись доступна по ссылке: http://iling-ran.ru/yanko/dialog/steklo3.wav).

Первые два предложения в составе сложного предложения *Потом я подошла в другую комнату* и *вот у меня выбито стекло* не завершают рассказа. Только падение на акцентоносителе конечной ремы *ощущаю* говорит об окончании повествования. В примере (1) имеются и другие значимые и незначимые (автоматические) изменения частоты основного тона, которые не релевантны для этого обсуждения и которые мы поэтому здесь не отмечаем.

В примере (2) представлена еще одна из стратегий маркирования незавершенности. Эта стратегия позволяет не только указать на незавершенность повествования, но и автономно выразить значения темы и ремы. Так, в примере (из корпуса [Spokencorpora 2017]), представляющем собой одно простое предложение, мы наблюдаем подъем на словоформе *сне*, маркирующий тему *во сне*, падение на акцентоносителе ремы *страха* и подъем незавершенности на конечной словоформе *охватило*:

(2) *Во* ***сне/*** *какое-то чувство* ***стра\ха охвати/ло…*** (http://iling-ran.ru/yanko/dialog/STRAX.wav)

Таким образом, пример (2) демонстрирует феномен совместимости внутри одного речевого акта как минимум трех коммуникативных значений, имеющих отдельные акцентоносители: темы, ремы и незавершенности.

---

[2] Акцентоносители незавершенности выделены здесь и в примерах ниже полужирным шрифтом, подъем частоты основного тона обозначается знаком / после гласного ударного слога словоформы-акцентоносителя, а падение — знаком \.

Возникает вопрос, какие типы речевых актов, кроме сообщений, совместимы с дискурсивной незавершенностью[3]. И если предположить, что показатели незавершенности присущи не только сообщению, но и другим типам речевых актов, то каковы параметры этих показателей?

Ответ на этот вопрос должен расширить наши представления о структуре и просодических средствах организации звучащего дискурса, а это, в свою очередь, послужит решению проблемы порождения и распознавания звучащей речи.

Для того, чтобы показать, что совместимость дискурсивной незавершенности с определенными типами речевых актов имеет место, достаточно привести пример или серию примеров подобной совместимости. Примеры не должны нарушать требований грамматической корректности, осмысленности и не быть контринтуитивными. В качестве источников для поиска примеров и анализа просодических показателей незавершенности используется просодически размеченный корпус «Рассказы о сновидениях и другие корпуса звучащей речи» [Spokencorpora 2017], Мультимедийный корпус МУРКО Национального корпуса русского языка (НКРЯ) и рабочий массив аудиозаписей, специально подготовленный автором для решения поставленной здесь задачи. Рабочий массив содержит записи 281 эпизода, которые включают различные типы речевых актов в контексте незавершенности, имеющей просодические средства выражения. Массив имеет общую продолжительность звучания около 1 часа 10 минут. Для анализа просодических показателей связности и других значений используется компьютерная система Praat [Boersma, Weenink 2012]. Звучащие записи примеров доступны на странице http://iling-ran.ru/yanko/dialog/.

Задача анализа дискурсивной незавершенности решается здесь для обращений (**Раздел 1**), императивов (**Раздел 2**) и вопросов (**Раздел 3**).

Для просодической разметки используются следующие знаки.

I. \ — падение частоты основного тона типа ИК-1 или ИК-2 с понижением на ударном слоге словоформы-акцентоносителя и дальнейшим понижением или ровным низким тоном на заударных слогах, если они есть ([Bryzgunova 1982: 97–122].

[3] Результаты, которые дают частичный ответ на вопрос о совместимости незавершенности с речевыми актами, отличными от сообщений, были получены в работе [Kobozeva 1999], где исследовалась «иллокутивная самостоятельность частей сложного предложения». Эта работа содержит примеры сложных предложений, имеющих в своем составе простые предложения с различной иллокутивной функцией. Соответственно, было показано, что совместимы императив и вопрос (*Дай мне почитать этот журнал, или ты сама его еще не прочла?*), а также сообщение и вопрос (*Ты, говорят, принес торт, так где же он?*). Между тем, если две иллокутивные функции совместимы в рамках одного сложносочиненного предложения, то можно сделать дальнейший вывод о том, что в пределах неконечного речевого акта следует ожидать какого-либо просодического показателя незавершенности. Таким образом, работа [Kobozeva 1999] показала, что некоторые иллокутивные функции совместимы в рамках одного сложного предложения, однако большинство более конкретных вопросов о природе такой совместимости тем не менее сохраняются.

II. / — подъем частоты основного тона на ударном слоге словоформы-акцентоносителя и падение на заударных слогах, если они есть. Если заударных в словоформе нет, заударное падение элиминируется (ИК-3, по [Bryzgunova 1982: 97–122]).

III. /– — подъем частоты на ударном слоге плюс ровные или слабо нисходящие заударные (ИК-6, по [Bryzgunova 1982: 97–122]).

IV. \/ — падение частоты основного тона или ровный низкий тон на ударном слоге акцентоносителя плюс подъем на заударных слогах если они есть. Если заударных слогов нет, интегральное нисходяще-восходящее движение тона фиксируется на конечном или единственном слоге словоформы-акцентоносителя (ИК-4, по [Bryzgunova 1982: 97–122][4].

1. Незавершенность и обращение

Обращения — это не один определенный тип речевых актов, а известное разнообразие типов обращений с различными вокативными иллокутивными функциями, см., например, [Zwicky 1974], [Yanko 2008: 98–106]. Различные функциональные типы обращений характеризуются различным просодическим оформлением. Обратимся к одному из типов обращений, которые свободно комбинируются с незавершенностью. Это обращения, направленные на привлечение внимания слушающего, с которым говорящий еще не вступил в контакт: *Зина, закрой окно*; *дорогой Вася, позволь поздравить тебя с получением Нобелевской премии!*; *здравствуйте, Иван Иванович!*. Если обращение состоит только из личного имени или имени деятеля, заменяющего имя, на ударном слоге имени фиксируется падение частоты основного тона (*Ва\ся!*; *Профе\ссор!*; *До\ктор!*; *Води\тель!*; *Исте\ц!*). Если обращение имеет в своем составе слова оценки *дорогой, уважаемый, милая* или слова вежливости (*господа, товарищи*): *дорогой Вася, господа офицеры, гражданин начальник, товарищ старший лейтенант* на первой словоформе обращения фиксируется подъем, на конечной — падение. Чтобы продемонстрировать сочетаемость таких обращений с незавершенностью, начнем с анализа обращения в отсутствие незавершенности. В дальнейшем будет продемонстрирован пример того же иллокутивного типа и в контексте незавершенности. Обратимся к фрагменту из кинофильма «Гараж». Герой В. Гафта обращается к герою Л. Маркова в нарочито официальном тоне:

(3) ***Многоуважа/-емый*** *Павел* ***Константи\ныч****! Вы крупный ученый, говорят, с мировым именем, вы член-корреспондент…*
(http://iling-ran.ru/yanko/dialog/Mnogouv.wav)

В примере (3) наблюдается подъем частоты тона на ударном слоге определения, затем следует относительно ровный (слабо нисходящий тон) вплоть до ударного слога конечной словоформы. В примере (3) это словоформа

[4] Нотация в основном продолжает традицию [Kodzasov, Bonch-Osmolovskaja, Zaharov, Kobozeva, Krivnova 2005, 2006].

Константиныч: ударный слог — *-ти-*. Обратившись к слушающему, говорящий останавливается, он не спешит с продолжением своей речи. Говорящий берет небольшую паузу и использует нисходящую трактовку конца обращения. Перед нами образец обращения, которое артикулируется как автономный и законченный шаг дискурса.

Между тем в примере (4) православный священник, называющий свою паству *дорогие братья и сестры*, избирает другую стратегию формирования стыка между обращением и последующим дискурсивным пространством. Говорящий артикулирует конечную словоформу обращения *сестры* с рельефным подъемом на ударном слоге словоформы *сестры*, за которым следует падение на заударном слоге. Перед нами один из восходящих акцентов. И он соответствует нашим ожиданиям относительно того, как должна выражаться незавершенность в обращении, непосредственно за которым следует другой речевой акт, продолжающий речь говорящего. В данном случае — это сообщение:

(4) *Дорогие братья и* ***се/стры****, сейчас мы начнем чтение вечерних молитв* (http://iling-ran.ru/yanko/dialog/Sestry.wav).

Аналогично, в примере (5) из воспоминаний ветерана войны обращение *товарищи офицеры* формируется говорящим как незавершенный шаг дискурса:

(5) *Това́рищи* ***офице́/ры*** */ сёдня бу́дем бра́ть Варша́ву* [НКРЯ][5] (http://iling-ran.ru/yanko/dialog/OficeryVarshava.wav).

На ударном слоге словоформы *офицеры* в обращении *товарищи офицеры* фиксируется подъем частоты основного тона.

Таким образом, мы показали, что обращения, направленные на привлечение внимания слушающего, могут формироваться как незавершенные шаги дискурса, если за ними следует какой-либо другой речевой акт, который служит продолжением речи того же говорящего. Между тем обращения, нацеленные на зов слушающего, находящегося на удалении от говорящего (*Вася-я!*), обращения, направленные на поддержание уже начатого общения (*Подвинься, Зин!*), обращения, сопровождаемые жестом «погрозить пальцем» (*Вася! Ты смотри у меня!*) в силу сложности прагматического контекста и уникальности средств просодического оформления (см. [Yanko 2008: 98–105]) опции указания на незавершенность дискурса не поддерживают. Незавершенность в композиции с обращением говорит о том, что за обращением следует другой речевой акт: сообщение, еще одно обращение, вопрос, императив.

[5] В примерах из НКРЯ мы сохраняем орфографию транскрипта, заменяя разбиение на сегменты соответствующими знаками препинания и добавляя знак акцента: в примере (5) это знак подъема частоты основного тона на словоформе *офицеры*.

2. Незавершенность и императив

Иллокутивная функция просьбы или команды имеет сегментные средства выражения. В русском языке это формы повелительного наклонения. Соответственно, просодия императива не служит средством выражения иллокутивной функции, а только формирует речевой акт, делая его автономным и отдельным от других речевых актов в потоке речи. Императив оформляется нисходящим движением частоты основного тона на словоформе-акцентоносителе. Акцентоноситель выбирается в соответствии с базовыми принципами выбора акцентоносителя в коммуникативно релевантном компоненте предложения [Yanko 2008: 38–52]. В примере (6) это собственно императив, на котором фиксируется нисходящий акцент типа ИК-2:

(6) *Ну́/* ***вспо́\мните****,* ***вспо́\мните****, ско́лько бы́ло весьма́ зло́бных, соверше́нно наро́дных стишко́в и пе́сенок на э́ту те́му* [НКРЯ] (http://iling-ran.ru/yanko/dialog/vspomnite.wav).

В примере (7) группа императива содержит дополнение, и в соответствии с принципами, изложенными в [Yanko 2008: 38–52], акцентоносителем группы императива становится словоформа *книгу*. Она несет нисходящий тон, говорящий о том, что речевой акт является конечным на соответствующем этапе дискурса. На глаголе *напиши* фиксируется предваряющий конечное падение подъем тона.

(7) *…а чё ты́ во́т про́сто та́к до́ма сиди́шь,* ***напиши****/ каку́ю-нибудь* ***кни́\гу*** [НКРЯ] (http://iling-ran.ru/yanko/dialog/NapishiKnigu.wav).

Перейдем к комбинации императива с дискурсивной незавершенностью. Пример (8) из диалога ведущей радиопередачи с гостем программы демонстрирует смену исходного нисходящего акцента императива на восходящий. Все шаги из приводимого фрагмента, кроме конечного, завершающего, шага, отмечены единообразной просодией незавершенности:

(8) *Сегодня в Москве проходит прощание с Арсением* ***Роги/нским****, я знаю, что вы сегодня там* ***бы/ли****, несколько слов* ***расскажи/те****, и, во-первых, и что там* ***бы/ло****, и вообще, вспомнить Арсения Рогинского, мне кажется, сегодня было бы* ***пра\вильно*** (http://iling-ran.ru/yanko/dialog/FelgSvanidzeImperativCut.wav).

Здесь подъем на императиве *расскажите* органично вписывается в серию подъемов на словоформах *Рогинским*, *были* и *было*, которые служат акцентоносителями обрамляющих императив сообщений. На акцентоносителе завершающего шага словоформе *правильно* фиксируется падение: показатели незавершенности отсутствуют, фрагмент дискурса подошел к концу.

В примере (9), где группа императива имеет более чем однословную структуру, акцентоносителями незавершенности служат словоформы *критерии* и *принципы*:

(9) *Поэ́тому когда́ мы́ обсужда́ем разли́чные аспе́кты междунаро́дной жи́зни,* ***дава́/йте*** *всё-таки име́ть каки́е-то* ***крите́/рии,*** *каки́е-то о́бщие* ***при́/нципы,*** *что́бы все́м бы́ло поня́тно, что в одни́х слу́чаях госуда́рственный перевоpóт непри́е́млем, и о́н до́лжен бы́ть непри́е́млем и в* ***други́\x*** *случаях* [НКРЯ].

После артикуляции словоформы *принципы*, несущей восходящий акцент незавершенности, говорящий задумывается, затем продолжает свою мысль и завершает шаг своего рассуждения падением на словоформе *других*.

Таким образом, примеры (8) и (9) в сравнении с примерами (6) и (7) демонстрируют совместимость незавершенности с речевым актом просьбы или команды. Кроме того, пример (8) свидетельствует о том, что императив свободно встраивается в ряд незавершенных речевых актов сообщений. Другие типы речевых актов, кроме сообщений, также могут служить «правым» контекстом для императива, несущего показатель незавершенности.

3. Незавершенность и вопрос

Начнем с краткого обзора просодии вопросительных предложений вне контекста незавершенности.

Базовые просодические модели для русских вопросительных предложений представлены примерами (10) и (11). Это вопрос с вопросительным словом и *да-нет*-вопрос. Базовые типы вопросов не отягощены дополнительными значениями, такими, как контраст, эмфаза, вызов, подчеркнутая вежливость, они не служат результатами линейно-акцентных преобразований или эллипсиса, они не содержат указаний на незавершенность дискурса. Вопрос (10) имеет подъем на вопросительном слове *как* и падение на конечной словоформе *чувствуете*:

(10) ***Ка́/к*** *вы́ себя́* ***чу́\вствуете***? [НКРЯ]
(http://iling-ran.ru/yanko/dialog/KakChuvstvuete.wav).

В *да-нет*-вопросе (11) имеется единственный акцентный пик на словоформе *летать*, это подъем частоты на ударном слоге словоформы типа ИК-3. На словоформе *можете* частота градуально снижается:

(11) *А вы́ и* ***лета́/ть*** *мо́жете*? [НКРЯ]
(http://iling-ran.ru/yanko/dialog/Letatj.wav).

Далее. *Да-нет*-вопросы с союзом *или* артикулируются особым образом. На акцентоносителе группы, которая представляет собой первый дизъюнктивный член, фиксируется подъем, а на акцентоносителе второй — падение, если дизъюнктивных членов два. Пример (12) содержит вопрос ведущей радиопрограммы о том, продолжит ли гость чтение стихов поэта Меркулова или перейдет к другому автору:

(12) *Вы будете продолжать щас Евгения* ***Мерку/лова*** *или найдете щас* ***еще*** *кого-то из вашей замечательной четверки*?
(http://iling-ran.ru/yanko/dialog/VoprosSili.wav).

Здесь мы наблюдаем подъем на словоформе *Меркулова* и падение на *еще*. Особенность подобного распределения акцентов в том, что если бы второй дизъюнктивный член был в предложении единственным, на нем фиксировалось бы не падение, а, наоборот, подъем: *Вы найдете* ***еще/*** *кого-то из вашей замечательной четверки*? Далее. Если дизъюнктивных членов более двух, падение фиксируется на акцентоносителе конечной группы, а на акцентоносителях неконечных групп фиксируются подъемы. В конъюнктивных же группах подъем фиксируется на акцентоносителе конечной группы, остальные конъюнктивные члены артикулируются безакцентно: *Это Вася, Коля и Константин/ с женой?*

При вынесении невопросительного компонента в начало вопроса — как вопроса с вопросительным словом, так и *да-нет*-вопроса — «темоподобное» начало принимает нисходящий акцент, ср. пример (13) с вопросительным словом:

(13) *А* ***цветы\ чьи****?* [НКРЯ] (http://iling-ran.ru/yanko/dialog/TsvetyChji.wav)

В вопросе (13) имеется два падения: на *цветы* и на *чьи*. Первое падение расположено на графике частот «выше» второго. Структура (12) получена путем линейно-акцентного преобразования базовой структуры *Чьи цветы*? с подъемом частоты на *чьи* и падением — на *цветы*.

Перейдем к анализу комбинаций различных типов вопросов с незавершенностью.

3.1. Незавершенность и вопросы с вопросительным словом

Пример (14) говорит о том, что при наложении на вопрос с вопросительным словом незавершенность имеет в качестве показателя ожидаемый подъем:

(14) *<… Не пособие какое-нибудь. Очень серьезное исследование. И здéсь вáжен экспериментáльный рядд. Скажúте,> почемý мы́ такáя закомплексóванная* ***нá/ция***? *<Почему у нас нет художественного понятия… слова, которое выражало бы сексуальные потребности человека!>* [НКРЯ] (http://iling-ran.ru/yanko/dialog/Natsija.wav).

В примере (14) перед нами два следующих друг за другом вопроса с *почему*, и первый из них имеет подъем тона на словоформе *нация*. Здесь перед нами не совсем каноническая конструкция типа ИК-3, а продолжение на заударном слоге подъема частоты, начавшегося на ударном слоге. В любом случае перед нами подъем тона, и наша гипотеза состоит в том, что этот подъем предваряет еще один вопрос, следующий за первым. Таким образом, пример (14) иллюстрирует способность вопроса с вопросительным словом вступать во взаимодействие с незавершенностью.

3.2. Незавершенность и *да-нет*-вопросы

Как было показано выше на примере (10), показателем *да-нет*-вопроса в русском языке служит подъем тона типа ИК-3. Следовательно, просодический маркер *да-нет*-вопроса фактически совпадает с одним из показателей

незавершенности. Наша гипотеза состоит в том, что отсутствие сочетаемости *да-нет*-вопросов с незавершенностью, которую мы наблюдаем в большинстве контекстов (об особых контекстах говорится ниже в подразделе 3.3), может объясняться именно совпадением показателей вопроса и незавершенности.

3.3. Незавершенность и серии вопросов

Сочетаемость вопросительных предложений с незавершенностью осложняется тем, что прозвучавший вопрос требует получения ответа от второго коммуниканта. Не ответить на вопрос или не дождаться ответа — значит решительно нарушить принципы кооперативной коммуникации. Следовательно, в дефолтной ситуации мы имеем дело с вопросо-ответной парой, в которой указание на незавершенность фрагмента дискурса для вопроса излишне: заданный вопрос иллокутивно «вынуждает» второго коммуниканта дать ответ [Krejdlin, Baranov 1992]. Следовательно, для постановки вопроса в контекст незавершенности необходимы специальные условия. Обратимся к особому типу вопросов, которые регулярно фигурируют в контексте незавершенности и которые именно в такой форме достаточно хорошо задокументированы в лингвистической литературе. Это серии вопросов, ср.: *Ваше имя? Возраст? Факультет? Курс?* [Bryzgunova 1982: 114], ср. также *Фамилия? Имя? Год рождения?* [Kobozeva 2005: 238]. Таким образом, для формирования ситуации незавершенности вопросу необходим контекст еще одного вопроса.

Вопросы в составе серий могут содержать начальное *А*, и в [Russian Grammar 1982: 390–390] такие вопросы называются вводящими: *А пространство? А время?*; *А учиться?*. [Russian Grammar 1982: 390–390] приписывает вводящим вопросам исключительно формы именительного падежа и инфинитива. Между тем, по нашим данным, вопросы такого типа не имеют ограничений на морфологическую форму группы, соединяющейся с *А*: *А в космос? А Васю? А побыстрее?* Такие вопросы представляют особенный интерес в контексте этой работы, т. к. они легко комбинируются с незавершенностью.

Вопросы анализируемого типа, как с *А*, так и без *А*, мы предлагаем трактовать как вопросы с эллипсисом, где собственно вопросительный компонент (вопросительное слово в вопросе с вопросительным словом или словоформа-акцентноситель в *да-нет*-вопросе) легко домысливаются из контекста: *Как ваше имя?* (ср. *Имя?*); *Какая у вас фамилия?* (ср. *Фамилия?*); *Какой факультет?* (*А факультет?*) *А летать* ***мо/жете***? (*А летать?*) *А ваш билет — где?* (*А ваш билет?*). Серии вопросов такого типа весьма характерны для допросов и опросов при заполнении анкет. Контекст незавершенности создают друг для друга именно вопросы, а не другие типы речевых актов, и говорящий просодическими средствами показывает, что он собирается задавать один вопрос за другим. При этом ответы на эти вопросы, даже если они в диалоге и имеются, не мешают намерению говорящего продолжить серию вопросов, которые он заготовил заранее. Говорящий дает понять слушающему, что текущий вопрос — не последний в ряду вопросов. Артикулируются такие вопросы с нисходящей или с нисходяще-восходящей просодией типа ИК-4. На последний

тип просодии в сериальных вопросах обратила внимание Е. А. Брызгунова [Bryzgunova 1982: 114]. В контексте серии такой вопрос может нести акцент ИК-4, и это говорит о том, что говорящий избирает стратегию создания серии вопросов. В отсутствие контекста незавершенности, а также при интерпретации говорящим вопроса в серии как независимого от контекста незавершенности, или как одиночного, эллиптические вопросы артикулируются с падением на словоформе-акцентоносителе. Пример (15) из кинофильма «Допрос» иллюстрирует ряд подобных вопросов, где соседствуют эллипитические вопросы и вопросы с вопросительным словом:

(15) — *Фамилия, имя-**о\тчество**? — Абиев Мурад Мехти-оглы.*
— *Год **рожде\/ния**? — Тыща девятьсот тридцать второй.*
— *Чем занимались, прежде чем стали начальником **це\/ха**? — Спортом.*
— *Как могло **случи/ться**, что вы оказались начальником галантерейного **це\ха**?*
(http://iling-ran.ru/yanko/dialog/KaljaginDopros2.wav)

В диалоге (15) первый вопрос артикулируется говорящим как находящийся вне серии (здесь говорящий еще не применил свою сериальную стратегию), затем второй и третий вопрос несут акцент типа ИК-4 с падением на ударном слоге и подъемом на заударных слогах, что говорит о включении в серию. При этом первый и второй вопрос имеют эллиптический характер, а третий и четвертый — это полноценные вопросы с вопросительным словом. Четвертый вопрос показателя продолжения серии уже не несет: он артикулируется в соответствии с базовым принципом артикуляции вопроса с вопросительным словом.

Анализ рабочего массива говорит о том, что сериальная стратегия представляет собой наиболее частотную и хорошо разработанную русской речью стратегию формирования незавершенности в контексте речевого акта вопроса. Серия создается из последовательности вопросов, подготовленных заранее, или позиционируемых говорящим как таковые. В серии объединяются вопросы с вопросительным словом и эллиптические вопросы. Полноценные (т. е. не-эллиптические) *да-нет*-вопросы могут быть включены в подобный ряд вопросов, но просодии, формирующей серию, они не принимают, сохраняя просодию автономного *да-нет*-вопроса. Это служит дополнительным свидетельством в пользу того, что *да-нет*-вопросы со средствами выражения дискурсивной незавершенности не взаимодействуют.

Итак, пример (15) говорит, во-первых, о том, что в контекст серии, могут быть включены как эллиптические вопросы, так и вопросы с вопросительным словом, где и те, и другие принимают соответствующую сериальному типу артикуляции нисходяще-восходящую просодию. И, во-вторых, пример (15) иллюстрирует совместимость в одном ряду вопросов, оформленных как сериальные, и вопросов, сохраняющих форму автономных. Кроме того, наши наблюдения говорят о том, что хорошие актеры, а следователя в фильме «Допрос» играет А. Калягин, в сериальных, перечислительных и других контекстах, основанных на повторе, склонны использовать разнообразные стратегии в одном и том же ряду единообразных речевых фигур.

***

Анализ массива данных говорит о том, что с дискурсивным смыслом незавершенности способны комбинироваться следующие из рассмотренных здесь типов речевых актов: 1) сообщение (это наилучшим образом изученный тип незавершенности), 2) обращения, нацеленные на привлечение внимания слушающего, с которым говорящий еще не вступил в коммуникацию, 3) императивы; 4) вопросы с вопросительным словом; 5) эллиптические вопросы с опущенным собственно вопросительным компонентом. По предварительным данным *да-нет*-вопросы в русской речи с показателями незавершенности не взаимодействуют, однако эта проблема еще требует дополнительного изучения. Мы предполагаем, что приведенный здесь анализ некоторых стратегий указания на то, что текущий речевой акт планируется говорящим как не последний в структуре монолога или диалога, не исчерпывает всего разнообразия стратегий незавершенности, разработанных русской речью. Мы надеялись лишь показать, что совместимость речевых актов, кроме сообщений, со значением незавершенности дискурса имеет место. Дальнейший анализ стратегий незавершенности — дело будущего.

References

1. *Bryzgunova E. A.* (1982) Intonation [Intonatsiya], Russian Grammar [Russkaya grammatika]. Vol. 1, Nauka, Moscow, pp. 98–118.
2. *Bryzgunova E. A.* (2003) Intonation and Syntax [Intonatsiya i sintaksis] // V. A. Beloshapkova (ed.). Modern Russian [Sovremennyj russkij jazyk]. Moscow: Azbukovnik, P. 869–903
3. *Boersma P., Weenink D.* (2012). Praat: Doing phonetics by computer. Version 5.3.04. Online: http://www.praat.org/.
4. *Kibrik A., Podlesskaja V.* (2009) Nightdream stories: A corpus-based study of spoken Russian discourse [Rasskazy o snovidenijah: Korpusnoe issledovanie ustnogo russkogo diskursa]. Moscow: Jazyki slavjanskih kul'tur.
5. *Kobozeva I. M.* (1999) On illocutionary independence of clauses in complex sentences [O kriterijah illokutivnoj samostojatel'nosti chastej slozhnogo predlozhenija] // Proceedings of the International Conference "Dialog99" [Trudy mezhdunarodnoj konferencii «Dialog 1999»]. P. 133–137.
6. *Kobozeva I. M.* (2005) An essay in characterizing lexical-syntactic, semantic, and pragmatic properties of interrogative dialogical turns in terms of features [Opyt razrabotki priznakovoj bazy dlja harakteristiki leksiko-sintaksicheskih, semanticheskih i pragmaticheskih svojstv voprositel'nyh replik] // Proceedings of the International Conference "Dialog 2005" [Trudy mezhdunarodnoj konferencii «Dialog 2005»]. P. 238–244.
7. *Kodzasov S. V., Bonch-Osmolovskaja A. A., Zaharov L. M., Kobozeva I. M., Krivnova O. F.* (2005) Data Base 'Intonation of Russian Dialogue: Interrogative Phrases' [Baza dannyh «Intonacija russkogo dialoga»: voprositel'nye repliki] // Proceedings of the International Conference "Dialog 2005". P. 245–247.

8. *Kodzasov S. V., Arhipov A. V., Bonch-Osmolovskaja A. A., Zaharov L. M., Krivnova O. F.* (2006) Data Base 'Intonation of Russian Dialogue: Commanding Propositions' [Baza dannyh «Intonacija russkogo dialoga»: pobuditel'nye repliki] // Proceedings of the International Conference «Dialog 2006». P. 236–242.
9. *Krejdlin G. E., Baranov A. N.* (1992) On illocutionary imposing in the structure of dialogue [Illokutivnoe vynuzhdenie v strukture dialoga] // Problems of linguistics [Voprosy jazykoznanija]. № 2. P. 84–99.
10. *Podlesskaya V.* (2016) "No po raschotu po moemu dolzhna rodit": the Russian Conjunction PO viewed through the prism of prosodically annotated corpus data // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2009), P. 574–579.
11. *Russian Grammar* (1982) [Russkaya grammatika]. Vol. 2. Moscow: Nauka.
12. *Spokencorpora 2017.* Prosodically Annotated Corpus of Spoken Russian. Pilot version. Online: http://spokencorpora.ru.
13. *Yanko T.* (2008) Intonational strategies of the Russian speech from a contrastive perspective [Intonatsionnye strategii russkoj rechi v soposnaviteljnom aspekte]. Moscow: Jazyki slavjanskih kul'tur.
14. *Yanko T. E.* (2017) Word Order and Accent Placement in Topics, Foci, and Markers of Discourse Continuity // Oslo Studies in Language. Comparative Slavic Syntax and Semantics. Vol 9, No 1. Pp. 45–57.
15. *Zwicky A.* (1974) Hey, whatsyourname! Chicago: Chicago Linguistics Society.

РУССКОЕ *КАК-НИБУДЬ* ПО ДАННЫМ ПАРАЛЛЕЛЬНЫХ КОРПУСОВ[1]

Зализняк Анна А. (anna.zalizniak@gmail.com)
Институт языкознания РАН; ФИЦ ИУ РАН, Москва, Россия

Денисова Г. В. (galina.denissova@unipi.it)
Пизанский государственный университет, Пиза, Италия

Микаэлян И. Л. (irina-mikaelian@yandex.ru)
Университет штата Пенсильвания, США

В докладе предлагается семантический анализ русского неопределенного наречия *как-нибудь*, проведенный на основе анализа данных французского, итальянского и английского параллельных подкорпусов НКРЯ, а также базы данных русских дискурсивных слов и их французских эквивалентов. В исследовании применяется унидирекциональный метод контрастивного анализа, при котором использованный профессиональным переводчиком способ передачи смысла анализируемой единицы текста оригинала рассматривается как ее квазитолкование, обнаруживающее возможные имплицитные компоненты ее значения. Проведенное исследование позволило подтвердить высокую степень лингвоспецифичности данного слова (обнаруживающую себя, с одной стороны, в значительной доле нулевых эквивалентов — как среди «моделей», так и среди «стимулов» перевода — а также в наличии широкого спектра различных «моделей» и «стимулов» перевода). При этом у слова *как-нибудь* было выявлено значение «маркера неконтролируемости», в ряде контекстов функционально сходное с конъюнктивом в романских языках, которое не зафиксировано толковыми и двуязычными словарями; с другой стороны, было обнаружено, что чисто оценочное значение ‘кое-как, плохо’ в современном языке значительно сузило свою сферу употребления по сравнению с 19-м веком и реализуется преимущественно одновременно с основным значением неопределенности образа действия.

Ключевые слова: русский язык, семантика, лексикография, параллельный корпус, неопределенные местоимения, неопределенные наречия, неконтролируемость, референция, снятая утвердительность

[1] Работа выполнена при поддержке РФФИ, грант №16-06-00339.

RUSSIAN *KAK-NIBUD'* THROUGH THE PRISM OF PARALLEL CORPORA

Zalizniak Anna. A. (anna.zalizniak@gmail.com)
Institute of Linguistics of the RAS; Institute of Informatics Problems of the FRC CSC RAS, Moscow, Russia

Denisova G. V. (galina.denissova@unipi.it)
Pisa State University, Italy

Mikaelian I. L. (irina-mikaelian@yandex.ru)
Pennsylvania State University, USA

The paper proposes a semantic analysis of the Russian indefinite adverb *kak-nibud'* based on the data collected from the French-Russian, Italian-Russian, and English-Russian parallel subcorpora of the Russian National Corpus, as well as from the Data Base of the Russian Discourse Markers and their French equivalents. The study applies the "unidirectional method" of contrastive analysis within which the translation by a professional translator is viewed as a quasi-lexicographic explication of a given unit revealing implicit components of its semantics. Our analysis demonstrates that *kak-nibud'* is a highly language-specific Russian word. It reflects in a high percentage of null equivalents of this unit in the three languages under investigation, for both Russian taken as the source or target language. The study has also allowed us to show that the analyzed adverb can function as a marker of non-controllability of a hypothetic event similar to the function of the subjunctive mood in Romance languages. On the other hand, the use of *kak-nibud'* ('anyhow', 'poorly') in a purely evaluative meaning cited by monolingual and bilingual dictionaries has shrunk in contemporary Russian compared to the Russian of the 19th century.

Key words: Russian language, semantics, lexicography, parallel corpus, indefinite pronouns, indefinite adverb, non-controllability, reference, non-veridicality

1. Водные замечания

Русское слово *как-нибудь*, с легкой руки В. И. Даля, процитировавшего пословицу *Русский крепок на трех сваях: авось, небось, да как-нибудь*, давно стало расхожим маркером «русского характера»[2]. Настолько, что даже в Интернете циркулируют сочинения, пытающиеся опровергнуть содержащуюся в этом

[2] См. обсуждение этого вопроса в Шмелев 2017.

слове апологию лени и наплевательства — характерным для любительской лингвистики способом восстановления его «истинного значения»[3].

Не имея в виду вступать в дискуссию о «русском характере», в данной статье мы проанализируем значение русского слова *как-нибудь*, основываясь, в том числе, на данных, предоставляемых параллельными подкорпусами НКРЯ и исходя из предположения, что использованный профессиональным переводчиком способ передачи смысла анализируемой единицы текста оригинала может рассматриваться как ее квазитолкование, обнаруживающее возможные имплицитные компоненты ее значения[4].

В словарях БАС и МАС у слова *как-нибудь* различаются три значения:

1. Каким бы то ни было образом, способом; так или иначе.

(1) — *Обо мне беспокоиться нечего. Меня друзья как-нибудь пристроят.* Салтыков-Щедрин, Пошехонская старина.

(2) *Арсений Романович торопился как-нибудь приладить подтяжку.* Федин, Необыкновенное лето.

2. *Разг*. Недостаточно хорошо, кое-как, небрежно[5].

(3) *Мы все учились понемногу Чему-нибудь и как-нибудь.* Пушкин, Евгений Онегин.

(4) *Посидели на пригорочке, Закусили как-нибудь (Не разъешься черствой корочки) И опять пустились в путь.* Н. Некрасов, Коробейники.

3. *Разг*. Когда-нибудь, как только найдется время.

(5) *В далеком лесу кричат грачи. Там на березах уйма гнезд. Мы с Ванькой уговорились как-нибудь сходить туда*. Замойский, Подпасок.

(6) — *Семен Михайлович! Заходи обязательно. Как-нибудь вечерком, часиков в семь.* А. Пришвин, Солнечная зима.

Будем называть первое значение значением образа действия (или основным), второе — оценочным, третье — временным.

---

[3] См., в частности, http://valhalla.ulver.com/f384/t17579.html.

[4] Этот принцип, обозначенный как «унидирекциональный метод контрастивного анализа», был предложен в Зализняк 2015.

[5] Семантический переход от значения неопределенности выбора к значению негативной оценки представлен также в наречии *кое-как*, для которого оценочное значение в современном языке является единственным. Ср. также франц. *n'importe quoi* 'что угодно' и *n'importe comment* 'как угодно', имеющие производные значения, соответственно, 'что попало, чепуха, чушь', 'как попало'.

Те же три значения выделяются в двуязычных словарях, где для них предлагаются следующие переводные эквиваленты:

1. франц. *d'une manière ou d'une autre, d'une façon ou d'une autre*; итал. *in qualche modo, in un modo o nell'altro*; англ. *somehow.*
2. франц. *d'une manière quelqonque, d'une façon quelqonque,* итал. *alla meglio, alla meno peggio, alla carlona*; англ. *anyhow.*
3. франц. *un jour ou l'autre*; итал. *un giorno*; англ. *some time.*

Однако если посмотреть на то, какие «модели перевода» реально используют переводчики при переводе с русского, а также, и даже в особенности, какие «стимулы перевода» вызывают появление интересующего нас слова в переводе на русский[6], то картина окажется существенно иной. А именно, бросаются в глаза, прежде всего, следующие три обстоятельства:

1) в значительной части случаев (около трети) русскому *как-нибудь* в переводе на французский, итальянский и английский языки не соответствует никакой лексемы — и, соответственно, в переводе *на* русский язык слово *как-нибудь* появляется как бы «ниоткуда»;
2) количество встречающихся типов межъязыковой эквивалентности («моделей перевода» и особенно «стимулов перевода») необычайно велико;
3) некоторая часть примеров употребления *как-нибудь* не укладывается ни в одно из перечисленных значений.

2. О статусе оценочного значения *как-нибудь*

Как известно, местоимения на *-нибудь* выражают экзистенциальную квантификацию, т. е. позволяют упоминать объект из некоторого класса, не индивидуализируя его, и относятся к категории нереферентных неопределенных местоимений (см. Падучева 2015, 2016 с дальнейш. библиогр.). К этой же категории, очевидно, следует отнести и слово *как-нибудь*, являющееся, соответственно, нереферентным неопределенным наречием. Главной особенностью данного класса слов является употребление исключительно в контексте снятой утвердительности (Падучева 1985: 94, 215–220; 2015, 2016). Местоимения на *-нибудь* не имеют конкретно-референтного употребления и поэтому недопустимы в контексте предиката, который описывает единичную ситуацию, имевшую место в прошлом, ср. **Он купил что-нибудь*, а также в контексте фактивного (**Хорошо, что он что-нибудь купил*) или импликативного (**Ему удалось что-нибудь купить*) подчиняющего предиката. Этот запрет снимается в контексте вопроса, побуждения, ирреального наклонения или нефактивного подчиняющего предиката (ср.: *Ты что-нибудь купил?*;

[6] Термины «модель перевода» и «стимул перевода» в том значении, в котором они здесь используются, были определены в [Loiseau et al. 2013].

Купи что-нибудь!; *Надеюсь, он что-нибудь купил* и т.д.), а также в некоторых других случаях, в частности, если слово на *-нибудь* находится в сфере действия квантора всеобщности (ср. *Он позвонил своим бывшим друзьям и каждому как-нибудь нахамил*).[7]

Для современного языка это ограничение касается всех значений слова *как-нибудь*: нельзя сказать **Меня друзья как-нибудь пристроили* (значение образа действия); **Мы переночевали как-нибудь* (оценочное); **Я к нему заходил как-нибудь на прошлой неделе* (временное). При этом в отношении оценочного значения эта норма изменилась по сравнению с 19-м веком: сейчас нельзя сказать ??*закусили как-нибудь* (ср. приводимый в МАС пример (4) из Некрасова).

Дело в том, что употребление слова *как-нибудь* в чисто оценочном значении по сравнению с 19-м веком существенно сузило сферу своего употребления: в современном языке *как-нибудь* возможно только в контекстах снятой утвердительности, где оценочное значение в той или иной степени соединено со значением образа действия (см. примеры (7)–(9) ниже). А за пределами этого типа контекстов оно было в значительной степени вытеснено другим бывшим неопределенным наречием — *кое-как*, имеющим близкое, хотя и не тождественное оценочное значение. Тем самым собственно оценочное значение у *как-нибудь* практически утрачено — оно может реализоваться лишь в контексте противопоставления (ср. пример (8)), в том числе — в конструкции [*не как-нибудь, а*] (см. ниже). В прочих случаях оценочный компонент в *как-нибудь* присутствует в форме коннотации у основного значения образа действия, ср. (7):

(7) Судьба уготовила мне быть старшей — маминой подручной. Все им, все им, младшим, сама уж ***как-нибудь***. [Нонна Мордюкова. Казачка (2005)]

В примере (8) оценочная коннотация повышается в ранге за счет обстоятельств, указанных в предтексте:

(8) И смертная казнь и пожизненное заключение одинаково безнравственны, но если бы мне предложили выбирать между казнью и пожизненным заключением, то, конечно, я выбрал бы второе. Жить ***как-нибудь*** лучше, чем никак. [А. П. Чехов. Пари (1888)]

В следующем примере из Гоголевского «Носа» (полностью соответствующего современной норме) оценочная коннотация дополнительно эксплицирована в последующем тексте, ср. пример (9) и его переводы на франц. язык[8]:

(9) Сделайте милость, — произнес Ковалев умоляющим голосом, — нет ли средства? ***как-нибудь*** *приставьте*; *хоть не хорошо, лишь бы только держался*; я даже могу его слегка подпирать рукою в опасных случаях. [Н. В. Гоголь. Нос (1832–1833)]

---

[7] См. полный список контекстов, лицензирующих употребление местоимений на *-нибудь* в Падучева 2015.

[8] В переводах мы выделяем полужирным курсивом фрагмент, соответствующий русскому *как-нибудь*, светлым курсивом — релевантный контекст. При помощи пометы ZERO мы маркируем нулевые модели и стимулы перевода.

(9-fr1) *remettez-le*, ***d'une façon ou d'une autre***; même pas bien, pourvu qu'il tienne;
(9-fr2) **ZERO** *Remettez-le en place*, même de travers, pourvu qu'il tienne
(9-fr3) *Arrangez*-le ***tant bien que mal***;
(9-fr4) *Recollez-moi* ***comme vous pourrez***. Même pas très bien, pourvu qu'il tienne.

Обратим внимание на то, что в (9-fr3) элементы русской фразы *как-нибудь* и *хоть нехорошо* переведены вместе одной единицей *tant bien que mal*, выражающей оценочный компонент значения русского *как-нибудь*, а компонент неопределенности способа действия и неполной контролируемости передан при помощи глагола *arranger*, являющегося эквивалентом для сочетания *как-нибудь* и *приставить*.

Чисто оценочное значение ('кое-как, не лучшим образом') в современном языке реализуется в одном из значений конструкции *не как-нибудь, а* — там, где эта конструкция выносит в фокус идею противопоставления по шкале «качества», ср. примеры (10)–(12):

(10) Она пела, играла на рояле, писала красками, лепила, участвовала в любительских спектаклях, но все это ***не как-нибудь, а*** с талантом [А. П. Чехов. Попрыгунья (1891)]

(10-en) She sang, she played the piano, she painted in oils, she carved, she took part in amateur performances; and all this ***not just anyhow***, but all with talent […] [Anton Chekhov. The Grasshopper (Constance Garnett, 1900–1930)]).

(11) Еврейской лошадкой был огромный чернобородый Лева Готлиб, которому удалось засунуть русскую Ирку в иудаизм, *да* ***не как-нибудь***, а по полной программе [Людмила Улицкая. Веселые похороны (1997)]

(11-it) Il cavallo ebraico era il gigantesco nero barbuto Lëva Gottlieb, che era riuscito a introdurre la russa Ira al giudaismo, *e* ***non per modo di dire***, ma secondo il programma completo [Ljudmila Ulickaja. Funeral party (Emanuela Guercetti)]

(12) Тем временем приближалась годовщина Сережиной смерти. Принять надо было человек тридцать, и ***не как-нибудь, а*** по-хорошему. [Людмила Улицкая. Зверь (1997)]

Ср. также пример (13-ru), где «стимулом перевода» для *как-нибудь* является идиома *trifle with fortune* 'играть с судьбой', выявляющая семантический компонент легкомысленного расчета на удачу и безответственного отказа от собственных усилий (тот самый, который участвует в поговорке, приводимой Далем).

(13) The thing failed this time, however, so the boys shouldered their tools and went away feeling that they had ***not trifled with fortune, but*** had fulfilled all the requirements that belong to the business of treasure-hunting. [Mark Twain. The Adventures of Tom Sawyer (1876)]

(13-ru) На этот раз им, однако, не повезло, и, взвалив на плечи лопаты, они ушли, сознавая, что отнеслись к делу ***не как-нибудь, а*** добросовестно

проделали все, что полагается искателям клада. [Марк Твен. Приключения Тома Сойера (Н. Дарузес, 1950)

Конструкция *не как-нибудь, а* может использоваться также и в ином, безоценочном, значении — как маркер неожиданности или экстраординарности сообщаемой информации, в том числе в контексте глаголов *звать*, *называть*, где *как-нибудь* соответствует обязательному актанту этих глаголов, ср. (14), (15):

(14) Можно представить себе, как местные мужички обрадовались десанту девушек, прибывшему на Колыму ***не как-нибудь, а*** по доброй воле. [Георгий Жженов. Прожитое (2002)]

(15) Он разговаривал с какой-то безликой дамой, — разумеется, на своем фрейбургском наречии, — все время называя ее ***не как-нибудь, а*** Оно. [Один абсолютно добрый волшебник (2004) // «Театральная жизнь», 2004.06.28]

3. Значение образа действия и идея неконтролируемости

Итак, основным значением *как-нибудь* является значение неопределенного образа действия. Анализ употребления этого слова и его эквивалентов в трех обсуждаемых языках (в обоих направлениях перевода) позволил выявить следующие основные типы контекстов, в которых реализуется данное значение.

1. Употребление в контексте глаголов со значением «преодоления» (*справиться, управиться*, *уладить*, *устроить*, *изловчиться, продержаться* и т. п., см. примеры (16)–(23)) — в качестве дополнительного маркера преодоления неполной контролируемости результата прилагаемых усилий. При этом сочетание такого глагола с *как-нибудь* указывает на возможность достижения этого результата — несмотря на дефицит средств и/или не важно каким способом («не важно как, но сделаю»).

(16) Зима уже кончается. Лето, короткое лето он ***как-нибудь*** *проживет*. Поташников остановился, ожидая Григорьева [Варлам Шаламов. Колымские рассказы, 1954–1962].

(16-it1) L'inverno stava finendo. Quanto all'estate, alla breve estate, ***l'avrebbe in qualche modo*** *superata*. Potašnikov si fermò per aspettare Grigor'ev [Varlam Shalamov. I racconti della Kolyma, Marco Binni, 1995].

(16-it2) L'inverno stava già finendo. E in estate, la breve estate di quei luoghi, ***ce l'avrebbe fatta a*** *sopravvivere*. Potašnikov si fermò ad aspettare Grigor'ev [Varlam Shalamov. I racconti di Kolyma, Sergio Rapetti, 1999].

Обратим внимание, что в переводе на итал. язык здесь появляется глагол *superare* 'преодолеть' или конструкция *farcela* <+ inf> со значением 'справиться'.

Во франц. переводе в подобных контекстах может появляться *tacher de* ‘стараться’, *arriver à* ‘получаться’, *s’arranger pour* ‘устроить так, чтобы’ и т. п., ср.:

(17) …тогда он встанет с постели на колена и начнет молиться жарко, усердно, умоляя небо *отвратить* ***как-нибудь*** угрожающую бурю. [Гончаров. Обломов]

(17-fr) alors il se levait de son lit, se mettait à genoux et commençait à prier avec ferveur, avec zèle, suppliant le ciel de ***s’arranger pour*** *lui épargner* la tempête qui le menaçait. [Ivan Gontcharov. Oblomov (Luba Jurgenson, 1988)]

В переводе на русский язык такое *как-нибудь* может появляться без какого-либо стимула в тексте-источнике, лишь в качестве «поддержки» предиката, обозначающего действие «с необеспеченным результатом», ср.:

(18) No, you go right along, Miss Mary Jane, and I’ll **ZERO** *fix it with* all of them. [Mark Twain. The Adventures of Huckleberry Finn (1884)]

(18-ru) Нет, вы уж поезжайте сейчас, мисс Мэри Джейн, а я тут с ними ***как-нибудь*** *улажу дело*. [Марк Твен. Приключения Гекльберри Финна (Н. Дарузес, 1950)]

(19) I must see what can **ZERO** *be done*. (Dickens, Charles / David Copperfield)

(19-ru) Подумаю, можно ли это ***как-нибудь*** *устроить*. (Диккенс, Чарльз / Жизнь Дэвида Копперфилда (www.lingvolive.com)

(20) Get good wages out there an’ we’ll put ‘em together. We’ll **ZERO** *make out*. [John Steinbeck. The Grapes of Wrath (1939)]

(20-ru) Что заработаем, пойдет в общий котел. ***Как-нибудь*** *выкарабкаемся*. [Джон Стейнбек. Гроздья гнева (Н. Волжина, 1940)]

(21) (21) “Let us hope to **ZERO** *hold* the fort till morning.” (Tolkien, John Ronald Reuel / The Fellowship of the Ring

(21-ru) — Может, до утра ***как-нибудь*** *продержимся*. (Толкиен, Джон Рональд Руэл / Братство Кольца) (www.lingvolive.com)

То же верно и для перевода с русского: поскольку *как-нибудь* является лишь дополнительным маркером семантического компонента преодоления неконтролируемости, заключенного уже в значении глагола, в целом перевод, содержащий эквивалент только для русского глагола, обозначающего действие с необеспеченным результатом, оказывается адекватным. Это может быть глагол с ингерентной семой неполной контролируемости результата; так, сочетание *как-нибудь* с глаголом *уладить*, может передаваться во французском одним глаголом *s’arranger*, ср.:

(22) Чтоб загладить свою вину, ты ***как-нибудь*** *уладь* с хозяином, чтоб мне не переезжать. [И. А. Гончаров. Обломов (1848–1859)]

(22-fr) Pour effacer ta faute, ***arrange-toi*** avec le propriétaire pour ne pas déménager. [Ivan Gontcharov. Oblomov (Luba Jurgenson, 1988)]

В других случаях, само *как-нибудь* вносит значение затрудненности, имплицирующей неполную контролируемость результата. В таком случае, использование в переводе неопределенного наречия или его аналога становится обязательным, ср.:

(23) — Ради того, чтобы ***как-нибудь*** *завязать разговор*, я временно примирюсь с вашим отказом. [В. В. Набоков. Ultima Thule (1940)]

(23-en) For the sake of ***somehow*** *starting our talk*, I shall temporarily accept your refusal. [Vladimir Nabokov. Ultima Thule (Dmitri Nabokov, Vladimir Nabokov, 1973)]

В английском языке наиболее частотным «стимулом перевода» для *как-нибудь* в сочетании с предикатом не полностью контролируемого действия оказывается глагол *to manage* 'справиться', часто (но не обязательно) в сопровождении наречия *somehow*.

(24) "I'll ***manage*** to *survive*." (Simak, Clifford D. / The Goblin Reservation)

(24-ru) И ***как-нибудь*** все это *переживу*. (Саймак, Клиффорд Д. / Заповедник гоблинов)

При этом само действие, с которым надлежит «справиться», в английском языке может быть не выраженным, а только подразумеваться:

(25) What'll become of country folks? Town folks can ***manage somehow***. They've always managed. [Margaret Mitchell. Gone with the Wind, Part 1 (1936)]

(25-ru) Какая участь ждет всех плантаторов? Городские жители ***как-нибудь*** *устроят* свою жизнь. Они всегда находили пути. [Маргарет Митчелл. Унесённые ветром, ч. 1 (Т. Озерская, 1982)]

2. Один из наиболее характерных контекстов употребления *как-нибудь* в значении «не важно как, но сделаю» — ответная реплика в диалоге. В таком контексте значение этого слова осциллирует, в семантической плоскости, между мужественным «будет трудно, но я справлюсь» и наплевательским «<пусть будет> любым способом, мне все равно», а в прагматической — между вежливым «не беспокойтесь за меня» и грубым «отстаньте от меня».

В примере (26) *как-нибудь* в ответной реплике выражает отказ от действия, предлагаемого собеседником:

(26) Раздевайтесь, пожалуйста, раздевайтесь! Николай Васильевич, по своему обыкновению, позволил себе избавиться от одной только шляпы, а затем сиротливо сложил руки на животе и сказал: — Да уж спасибо, спасибо… ладно… ***Я уж так как-нибудь***. [Андрей Волос. Недвижимость (2000) // «Новый Мир», 2001]

Выражаемое таким образом желание отстранить собеседника от решения своих проблем может быть как актом вежливости, так и грубости, ср. примеры (27), (28) и (29):

(27) — Серьёзные? — Д-да… Серьёзные. Но ты не морочь себе голову… я разберусь… ***как-нибудь***. Мне надо подумать… хорошо подумать. Как следует. Ты меня извини, милый, я пойду к себе. [Вера Белоусова. Второй выстрел (2000)]

В русском переводе такое *как-нибудь* может появляться «ниоткуда» — как средство передачи коммуникативного намерения освободить собеседника от заботы о себе или, наоборот, отстранить его от решения своих проблем:

(28) You have been more than kind. I can *show* **ZERO** *myself around*." [Dan Brown. The Da Vinci Code (2003)]

(28-ru) Вы и без того потратили на меня время, сестра. Дальше я ***как-нибудь*** сам. [Дэн Браун. Код Да Винчи (Н. Рейн, 2004)]

(29) Я же мать. И хочу тебе счастья. — У нас теперь много других поводов для беспокойства. Я со своей личной жизнью ***как-нибудь*** сама разберусь. — Какая у тебя личная жизнь? — Все, мам, спокойной ночи. [Маша Трауб. Замочная скважина (2012)]

Особо следует отметить изолированное употребление *как-нибудь*, где это слово выражает идею «справлюсь без вашей помощи» само по себе, оно не соотносится ни с каким предикатом в предтексте, ср.:

(30) Ах, что вы говорите — дальний путь.
Какой-нибудь ближайший полустанок.
Ах, нет, не беспокойтесь. ***Как-нибудь***.
Я вовсе налегке. Без чемоданов.
[Бродский. Мне говорят, что нужно уезжать...][9]

(31) А уютный подвальчик, черт меня возьми! Один только вопрос возникает, чего в нем делать, в этом подвальчике? […] — Зачем вы меня тревожите, Азазелло? — спросила Маргарита, — ***как-нибудь***! [М. А. Булгаков. Мастер и Маргарита (1929–1940)

(31-it) — Una cantina simpatica, diavolo! Ci si domanda una cosa soltanto: che cosa fare in questa cantina? — […] — Perché m'inquieta, Azazello? — chiese Margherita. ***In qualche modo ci si arrangia*** [Mikhail Bulgakov. Il Maestro e Margherita, Vera Dridso, 1967]

(31-en) A cosy little basement, devil take me! Only one question arises — what is there to do in this little basement? […] "Why do you trouble me, Azazello?" asked Margarita. ***"We'll live somehow or other!"*** Mikhail Bulgakov. Master and Margarita (Richard Pevear, Larissa Volokhonsky, 1979)]

---

[9] По-видимому, наиболее адекватным переводом этого примера на английский язык было бы *I will manage*. Как уже отмечалось выше, глагол *to manage* способен выражать идею 'справиться' безотносительно к конкретному предикату.

Для примера (31) в итальянском и английском переводах восстанавливается подразумеваемый предикат общей семантики.

3. Существует еще один тип употребления *как-нибудь*, который не зафиксирован словарями, но который, по-видимому, следует считать отдельным значением этого слова — это употребление в гипотетическом контексте, когда речь идет о нежелательном событии, которое нужно предотвратить, но которое не контролируемо для говорящего. Оно включает элемент неопределенности способа осуществления действия и неопределенности временного момента, к которому оно приурочено — но при этом не может быть сведено ни тому, ни к другому. Так, в примере (32) речь идет, очевидно, не о том, что ребенок может захныкать тем или иным образом, и не о том, что это может произойти в тот или иной момент, а именно о самой возможности наступления этого негативного события и желательности его предотвращения (ср. также (33)). Это значение реализуется прежде всего в составе конструкции [*чтобы* (*как-нибудь*) *не*] которая может быть подчинена глаголу, выражающему опасение (примеры (34) и (35)); в отсутствии таких глаголов, именно наречие *как-нибудь* само вносит значение опасения (примеры (32), (32) и (36)). При этом в переводе чаще всего этот смысловой компонент неконтролируемости, вносимый словом *как-нибудь*, не передается никакими лексическими средствами.

(32) Она, кажется, унимала его, что-то шептала ему, всячески сдерживала, чтоб он ***как-нибудь*** *опять* не *захныкал* [Ф. М. Достоевский. Преступление и наказание (1866)]

(32-fr) — Elle cherchait, semblait-il, à l'apaiser, lui chuchotait quelque chose, s'efforçait de *l'empêcher* **ZERO** de se remettre à pleurnicher (Élisabeth Guertik, 1947)]

(32-it) — Sembrava che cercasse di calmarlo, gli sussurrava qualcosa, o distraeva in tutti i modi perché **ZERO** non *si rimettesse* a frignare [Fedor Dostoevskiy Delitto e castigo, Giorgio Kraiski, 1969].

(33) Он почел нужным предуведомить об этом сына, чтобы тот ***как-нибудь*** *не рассердился* [И. С. Тургенев Отцы и дети, 1860–1861]

(33-it) — Credette necessario prevenirne il figlio, acciocchè questi non **ZERO** *avesse poi ad andare in collera* [Ivan Turgenev Padri e figli, Federico Verdinois, 2016].

(34) Боялся я ужасно, чтоб меня ***как-нибудь*** *не увидали, не встретили, не узнали*.

(34-en) I was fearfully afraid of being **ZERO** seen, of being met, of being recognised. [Dostoevsky, Fyodor / Notes from the Underground] (www.lingvolive.com)

(35) Ее присутствие доставляло мне удовольствие, какого я уже давно не испытывал, и я боялся смотреть на нее, чтобы мой взгляд ***как-нибудь*** *не выдал* моего скрытого чувства. [А. П. Чехов. Жена (1891)]

(35-en) Her presence gave me a pleasure such as I had not felt for a long time, and I was afraid to look at her ***for fear my eyes would betray*** my secret feeling. [Anton Chekhov. The Wife (Constance Garnett, 1900–1930)]

(36) Не ходите, голубчик! Еще ***как-нибудь*** *попадетесь*! Не надо! — посоветовал Николай. [Максим Горький. Мать (1906)]

(36-en) “Don’t go, darling! ***Maybe*** you’ll get caught. You mustn’t!” Nikolay advised. [Maxime Gorky. Mother (D. J. Hogarth, 1921)]

Заметим, что в итальянском и французском языках глагол, находящийся в сфере действия модального оператора, выражающего опасение, по правилам этих языков стоит в форме конъюнктива, который маркирует значение неопределенности, нереферентности и, как следствие, неконтролируемости гипотетического события. В подобных контекстах русское *как-нибудь* оказывается функциональным эквивалентом конъюнктива в романских языках (ср. примеры (32-it) и (33-it)). Что касается английского языка, то в нем это значение может передаваться дискурсивным модальным маркером *maybe*, и при этом значение опасения утрачивается (36-en), или более сложной конъюнктивной конструкцией, эксплицитно вводящий лексему со значением опасения, как в примерах (35-en) и (37-en) (*for fear my eyes would betray*, *for fear I might*). Ср. также пример (39), где конструкция [чтобы не], включающая *как-нибудь* возникает в переводе с английского на русский.

Хотя чаще всего обсуждаемое употребление встречается в контекстах, относящихся к ситуациям, внешним для говорящего, оно также возможно в отношении неконтролируемых собственных действий[10], ср. (37) и (38):

(37) Не прикажете ли, я на крылечке постою-с... чтобы ***как-нибудь*** невзначай чего не подслушать... потому что комнатки крошечные.

(37-en) “Wouldn’t you like me to stand on the steps . . . ***for fear I might*** by chance overhear something . . . for the rooms are small?” Dostoevsky, Fyodor / The possessed

(38) С чувством неизъяснимого страха бросился он к столу, придвинул зеркало, *чтобы* ***как-нибудь*** *не поставить нос криво*. [Н. В. Гоголь. Нос (1832–1833)]

(38-fr) Empli d’un sentiment de frayeur indicible, il se précipita vers la table, avança le miroir, *pour ne pas* ***risquer de*** *remettre* son nez *de travers*. [Nikolaï Gogol. Le Nez (André Markowicz, 2007]

(39) He now kept carefully out of reach of ***any possible*** splash as Pedro and Phil snorted and fooled in their foul bath. [Vladimir Nabokov. Ada, or Ardor (1968)]

(39-ru) Теперь он следил за тем, чтобы Педро и Фил, которые, всхрапывая, бултыхались в нечистой купальне, ***как-нибудь*** его не обрызгали. [Владимир Набоков. Ада, или Радости страсти (С. Ильин, 1996)]

[10] Ср. Зализняк, Левонтина 1996.

Во французском переводе в этом случае возможно наречие со значением ‘случайно’, глагол со значением ‘рисковать’ и другие единицы, включающие вероятностную составляющую (ср. (38-fr)).

В заключение приведем несколько более современных примеров из основного корпуса НКРЯ.

(40) Официант, стоявший за спиной Абесаломона Нартовича, принес стул, и меня усадили рядом с товарищем из министерства, и я почувствовал, как он с ненавистью сжался, чтобы не притрагиваться ко мне, и я сам сжался, *чтобы* ***как-нибудь*** *не* прикоснуться к нему. [Фазиль Искандер. Сандро из Чегема (Книга 2) (1989)]

(41) А «византийцы» если и вспомнят по возвращении из волжской прогулки об этом «Н[овом] М[ире]», то лишь с тем, чтобы ***как-нибудь*** не завязаться в этом деле. — [А. Т. Твардовский. Рабочие тетради 60-х годов (1968) // «Знамя», 2003]

(42) Больше ей незачем стало сюда приходить, и даже напротив — у нее возникло чувство, будто следует держаться подальше, *чтобы* ***как-нибудь*** *не* изуродоваться, не поддаться чуждым и враждебным превращениям. [Ольга Славникова. Стрекоза, увеличенная до размеров собаки (1995–1999)]

4. Временно́е значение

По данным корпусов употребление *как-нибудь* во временно́м значении составляет около трети всех его вхождений. В словаре МАС это значение толкуется как разговорное: «когда-нибудь, когда найдется время». К этому можно добавить еще «когда обстоятельства сложатся благоприятствующим образом». В действительности, как семантические, так и прагматические различия между *когда-нибудь* и *как-нибудь* столь велики, что замена одного на другое оказывается практически никогда невозможна. Как и в значении образа действия, говорящий может использовать временное *как-нибудь* для небрежной отговорки, для вежливого отказа или для необязательного приглашения.

(43) Под конец скучного и незатянувшегося вечера Соня поняла — у нее нет ничего общего ни с Ирой, ни с Антониной. Нет, они вспомнили о Кирюше, Наташе и Вячеславе. Рассказали гостям, как оказались в одном доме. Разъезжаясь, договорились созвониться «как-нибудь». [Маша Трауб. Домик на Юге (2009)]

Здесь кавычки, маркирующие цитатность употребления, одновременно подчеркивают необязательность обещания сделать что-нибудь *как-нибудь*, которая представляет собой прагматическое приращение. *Как-нибудь* более вежливое, чем *когда-нибудь*, в котором неопределенность времени закреплена в значении, плюс имеется коннотация ‘нескоро’.

(44) Объясню вам ***как-нибудь*** *в другой раз*, — ответила Зина, всё не спуская с него взгляда. [В. В. Набоков. Дар (1935–1937)]

(44-en) I'll explain it ***some other time***, replied Zina, not taking her eyes off him. [Vladimir Nabokov. The Gift (Michael Scammel, Vladimir Nabokov, 1962)]

(45) Call me ***sometime***, Lily. You were great last night." [Lauren Weisberger. The Devil Wears Prada (2003)]

(45-ru) Лили, звякни ***мне как-нибудь***, ты в постели что надо. [Лорен Вайсбергер. Дьявол носит Прада (М. Маяков, Т. Шабаева, 2006)]

Отметим, что во временном значении *как-нибудь* может иметь особую фонетическую реализацию, в которой оно приближается к частице ([какнть]).

5. Заключение

Проведенный анализ позволяет представить значение русского неопределенного наречия *как-нибудь* следующим образом.

В своем основном значении *как-нибудь* указывает на то, что говорящий/персонаж собирается достичь результата, несмотря на недостаток средств, не важно каким образом; при этом предполагаемый результат обычно характеризуется как не самого высшего качества. В составе ответной реплики диалога в высказывании от 1-го л. *как-нибудь* выражает также желание говорящего вежливо оградить (или, наоборот, грубо отстранить) собеседника от своих проблем.

Идея неполной контролируемости, имплицируемая сочетанием неопределенности способа и времени осуществления действия, может реализоваться также в форме 'говорящий/персонаж опасается, что произойдет нежелательное и неконтролируемое им событие'.

Наконец, слово *как-нибудь* может указывать на то, что обсуждаемое действие говорящего произойдет в неопределенном будущем и при благоприятствующих обстоятельствах. Оно используется преимущественно в диалоге — в том числе, в качестве неопределенного обещания или даже отговорки.

Литература

1. *Зализняк Анна А.* (2015), Лингвоспецифичные единицы русского языка в свете контрастивного корпусного анализа // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог'2015. С. 651–662.
2. *Зализняк Анна А., Левонтина И. Б.* (1996), Отражение национального характера в лексике русского языка (размышления по поводу книги: Anna Wierzbicka. Semantics, Culture, and Cognition. Universal Human Concepts in Culture-Specific Configurations. N.Y., Oxford, Oxford Univ. Press, 1992) // Russian Linguistics, vol. 20, 1996, pp. 237–264.
3. *Падучева Е. В.* (1985), Высказывание и его соотнесенность с действительностью. М.: Наука. 1985.

4. *Падучева Е. В.* (2015), Нереферентные местоимения (на *-нибудь*) // Материалы для проекта корпусного описания русской грамматики (http://rusgram.ru). На правах рукописи. М., 2015.
5. *Падучева Е. В.* (2106), Местоимения типа что-нибудь в отрицательном предложении // Вопросы языкознания, №3, 2016. С. 22–36.
6. *Шмелев А. Д.* (2017), Русские авось и небось revisited // Die Welt der Slaven, Jg 62/2, S. 276–303.
7. *Loiseau S., Sitchinava D. V., Zalizniak Anna A., Zatsman I. M.* (2013), Information technologies for creating the database of equivalent verbal forms in the Russian-French multivariant parallel corpus // Информатика и ее применения, 2013. Том 7, вып. 2. С. 100–109.

ДВА ДИАЛЕКТА РУССКОЙ ГРАММАТИКИ: КОРПУСНЫЕ ДАННЫЕ И МОДЕЛЬ[1]

Циммерлинг А. В. (fagraey64@hotmail.com)
ГИРЯ им. А. С. Пушкина, МПГУ,
Институт языкознания РАН, Москва, Россия

TWO DIALECTS OF RUSSIAN GRAMMAR: CORPUS DATA AND FORMAL MODELS

Zimmerling A. V. (fagraey64@hotmail.com)
Pushkin State Russian Language Institute, Moscow State University of Education, Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia

This paper is addressed the problem of parametric variation in Russian grammar, with focus on copular constructions with agreeing and non-agreeing adjectival predicates. Basing on Russian National Corpus, I reconstruct two dialects of Russian morphosyntax. They differ regarding the assignment of the predicative instrumental case, raising conditions and the distribution of agreeing vs non-agreeing predicates after *быть* 'be', *стать* 'become' and *казаться* 'seem'. Russian-A only licenses predicative instrumental on adjectives after SEEM (*казалось странным, что P*) and non-agreeing predicatives after non-zero forms of BE or BECOME (*было странно, что P*). Russian-B allows non-agreeing forms after SEEM (*казалось странно, что P*) and forms of the predicative instrumental case after non-zero forms of BE and BECOME (*было странным, что P*). I argue that the differences between Russian-A and Russian-B must explained in terms of parametric settings and claim that Russian predicatives lack forms of the predicative instrumental. The assignment of the predicative instrumental to adjectival heads can be explained as subject control in all dialects, but only Russian-B allows raising of sententional arguments to the position of the matrix subject.

Keywords: Russian, corpus linguistics, parametric grammar, adjectives, predicatives, predicative instrumental case, agreement, control, raising, dialect variation

[1] Работа выполнена при поддержке гранта РНФ 16-18-00203а «Структура значения и ее отображение в системе лексических и функциональных категорий русского языка», реализуемого в МПГУ. Я благодарю анонимных рецензентов за высказанные замечания.

0. Структура исследования

В разделе 1 характеризуется параметрический подход к описанию внутриязыкового варьирования и вводится понятие диалекта грамматики. В разделе 2 обсуждаются параметры, связанные с подъемом аргумента и выдвигается гипотеза о том, что приписывание творительного предикативного (далее — ТП) во всех диалектах русской грамматики является случаем подлежащного контроля. В разделе 3 выделяются рестриктивный и нерестриктивный диалекты, отличающиеся условиями подъема. В разделе 4 обобщается дистрибуция форм ТП и несогласуемого предикатива по данным НКРЯ. В разделе 5 уточняются параметры русской грамматики.

1. Варьирование параметров грамматики: постановка проблемы

Для большинства работ в области русистики характерно представление о единстве литературной нормы[2]. Варьирование ограничительных условий на воспроизводство грамматических структур в текстах XX–XXI вв. обычно описывается в терминах «правильно» (соответствие эталонному варианту русского языка, предположительно являющемуся единым для всех, овладевших литературной нормой) vs «неправильно» (отклонение от эталона). Применение подобного подхода сопряжено с тремя проблемами. Во-первых, многие ограничительные условия слабо рефлектируются: носители языка сталкиваются с ними при анализе отрицательного языкового материала, т. е. при исправлении ошибок или в лингвистическом эксперименте. Во-вторых, нормативные описания русского словаря и грамматики [Ожегов 1964; Шведова 1982] ориентированы на тексты художественной литературы XIX — первой половины XX вв., степень соответствия которых современному узусу неясна. В третьих, не раскрыты механизмы, порождающие предполагаемые ошибки, и возникает иллюзия, что грамматика держится только на внешнем принуждении и без сознательного освоения в процессе обучения эталонной форме родного языка не функционирует.

Структуралисты представляют разговорный русский язык как автономную систему, сосуществующую с литературным языком [Красильникова 1982; Лаптева 1976], что является шагом вперед в осмыслении первых двух проблем, но не дает ориентира для решения третьей — неясно, как систему правил языка А (разговорный язык) вывести из системы правил языка В (литературный язык), и наоборот. Обзорные социолингвистические работы предлагают инвентаризацию «сильных» (менее подверженных варьированию) и «слабых» зон, например, устойчивых или неустойчивых позиций граммем ср. р. и дат. п. в структуре категорий рода и падежа существительных [Comrie, Stone, Polinsky 1996; Гловинская 1998].

---

[2] Здесь и далее понятие «норма» понимается в значении «внутренняя норма», т. е. условие грамматической правильности, присущее языковой подсистеме или набору параметров и действующее независимо от внешнего принуждения.

Более гибким представляется подход, при котором внутри- и межязыковое варьирование объясняется настройками микропараметров. Смысл параметризации в том, чтобы с определенным значением параметра был связан не один признак (в таком случае параметр не отличается от конкретно-языкового правила), а группа свойств, образующих кластер. Тем самым, параметризация одновременно исчисляет языковое разнообразие и ограничивает его [Лютикова, Циммерлинг, Коношенко 2016]. В работах, посвященных внутриязыковому варьированию, формы языка, различающиеся значениями общих параметров, часто называют «диалектами» [Harris 1996; Henry 1998; Wilson, Henry 1998]. Мы будем далее использовать этот термин, сознавая его минус — наличие омонимичного употребления в диалектологии, где оно указывает на территориальную форму языка, связанную с древним диалектным членением в ареале, где может быть параллельно представлена кодифицированная наддиалектная форма того же языка. Никаких утверждений о локализации форм русского языка, различающихся настройками обсуждаемых ниже параметров, мы не делаем.

2. Предикативный падеж и предикативный атрибут

2.1. Фонологическая выраженность глагольной вершины

В ряде славянских языков, ср. польский, ТП может приписываться вершине именного сказуемого нулевой связкой «быть». В современном русском это запрещено: ТП приписывается именной (1а) и адъективной (1б) вершине исключительно ненулевой формой глагола [Matushansky 2008]; [Bailyn 2012: 194], ср. невозможность (1в–г), хотя реликты прежнего состояния есть еще у А. С. Грибоедова, ср. (2).

(1) а. Она была/считалась известной.
б. Она была/считалась известной певицей.
в. *Она известной певицей.
г. *Она известной.

(2) А тетушка? Все девушкой, Минервой? Все фрейлиной Екатерины Первой? [А. С. Грибоедов. Горе от ума (1824)]

Исключения единичны, ср. клише *X молодцом,* представленное в НКРЯ 55 раз. Попытка вставить в контекст <*X — Z-ом*> другой оценочный предикат приводит к аномалии:

(3) *Он сволочью.

2.2. Ненулевая форма подлежащего

Другое условие употребления ТП связано с наличием внешне выраженного контролера согласовательной формы рода и числа прилагательного.

Параллелизм поведения предикативного атрибута в случае, когда контролером выступает подлежащее финитной клаузы, ср. (4а), и в случае, когда контролером служит дополнение, одновременно выступающее в роли субъекта т. н. малой клаузы, т. е. нефинитной предикативной группы, ср. (4б), побуждает считать малые клаузы типа (4б) трансформами (4а). Ненулевой подлежащный контролер рода и числа в стандартном случае стоит в той же клаузе, что предикативный атрибут[3].

(4) а. Катя$_{\text{NOM.SG.F}}$ была уже здоровой$_{\text{ADJ.NOM.SG.F}}$.
б. Вася$_{\text{NOM.SG.M}}$ застал [$_{\text{SC}}$ Катю$_{\text{ACC.SG.F}}$ уже здоровой$_{\text{ADJ.NOM.SG.F}}$].

Правила приписывания ТП, иллюстрируемые (1)–(4), обобщает условие (i).

(i) ТП в русском языке приписывается ненулевой формой финитного глагола, подлежащее которого контролирует форму предикативного атрибута в роде и числе.

2.3. Неканонические подлежащие и контроль

В русском языке есть несколько типов неканонических подлежащих — дативные аргументы независимых инфинитивных предложений (*Грузовикам здесь не проехать*) и несогласуемых предикативов (*Мне стыдно, грустно*), нулевые неопределенные местоимения 3 л. с семантикой агенса (∅$^{\text{3SG}}$ *Улицу засыпало песком*, ∅$^{\text{3PL}}$ *Улицу засыпали песком*), сентенциальные аргументы ([$_{\text{InfP}}$*Признаваться в ошибке*] *стыдно, Грустно,* [$_{\text{CP}}$ *что они уехали*]) и эксплетивное местоимение *это* (*Это было грустно*) [Zimmerling 2009]; [Циммерлинг 2012]; [Летучий 2014]. Большинство неканонических подлежащих не контролируют согласование. Исключение составляют эксплетив *это* и сентенциальные аргументы, хотя лексически заданной модели согласования здесь нет — предикативный атрибут стоит в форме ед. ч. ср. р., а ненулевая связка — в 3л. ед. ч. ср. р., т. е. реализуется т. н. дефолтное согласование. ТП возможен ровно в том случае, когда позиция подлежащего заполнена ненулевым выражением, поэтому приписывание ТП — одно из проявлений подлежащного контроля.

2.4. Эксплетив *это* как контролер предикативного атрибута

В предложениях с эксплетивом *это* ТП появляется при ненулевой форме глаголов *быть*, *стать*, см. (5в) и *казаться*, см. (5д). При *быть* и *стать* форма ТП конкурирует с исходной формой предикативного атрибута на *–о* (далее — ИФ) — *это было неумным/неумно*, см. (5г). ИФ в структуре без лексически заданного согласования можно рассматривать либо как им.-вин. п. ср. р. ед. ч. краткого прилагательного [Попов 1881], либо как омонимичную ей форму несогласуемого предикатива [Щерба 2008]. Примем вторую точку зрения и будем

[3] В данной статье мы не обсуждаем контроль согласования через границу клаузы и примеры типа ??*Вася$_i$ застал Катю$_j$ здоровым$_i$*.

далее обозначать предикатив пометой PRED. С такой разметкой дистрибуция (5а-е) непосредственно отражает распределение вариантов с контролируемой формой предикативного атрибута и неконтролируемой формой предикатива.

(5) а. Это неумно$_{\text{PRED}}$.
б. *Это неумным$_{\text{INSTR.SG.N}}$.
в. Это было н е у м н ы м $_{\text{INSTR.SG.N}}$.
г. Это было неумно$_{\text{PRED}}$.
д. Это кажется мне н е у м н ы м $_{\text{INSTR.SG.N}}$.
е. ?Это кажется мне неумно$_{\text{PRED}}$.

По условию (i), при нулевой связке возможна только ИФ. При ненулевой связке *быть*, а также полузнаменательных глаголах *стать, являться, делаться* и т. п., ТП и ИФ конкурируют. При *казаться*, *показаться*, *представиться* формальные описания русского языка объявляют форму ТП безальтернативной [Bailyn 2012: 196], а о варианте (5е) не упоминают. Но корпусные данные показывают, что в одном из диалектов русской грамматики ИФ сочетается с *казаться* и *показаться*:

(6) Это показалось мне с а н т и м е н т а л ь н о , п р и т о р н о и неумно, а между тем я находился уже в таком настроении, когда во всем искал прежде всего «глубины мысли». [А. П. Чехов. Огни (1888)]

2.5. Подъем эксплетива

Глаголы *казаться*, *показаться* являются операторами подъема, преобразование зависимой финитной предикации в нефинитную сопряжено с продвижением подлежащего зависимой клаузы в главную клаузу. (5д), повторенное ниже как (7б) — результат преобразования (7а). Вариант (7в) блокируется условием (i), так как в зависимой клаузе нет ненулевой финитной вершины.

(7) а. Мне кажется, [$_{\text{CP}}$ что *это* неумно$_{\text{PRED}}$].
б. *Это* кажется мне неумным$_{\text{INSTR.SG.N}}$.
в. *Мне кажется, что *это* неумным.

Эксплетив *это* в (7) ведет себя так же, как обычная именная группа (ИГ), контролирующая согласование в роде и числе, см. (8а)–(8в).

(8) а. Мне кажется, [$_{\text{CP}}$ что [$_{\text{DP}}$ это решение]$_{\text{NOM.SG.N}}$ неумно$_{\text{ADJ.NOM.SG.N}}$].
б. [$_{\text{DP}}$ это решение]$_{\text{NOM.SG.N}}$ кажется мне неумным$_{\text{INSTR.SG.N}}$.
в. *Мне кажется, что [$_{\text{DP}}$ это решение]$_{\text{NOM.SG.N}}$ неумным.

Замена ИФ ⇒ ТП сопряжена с продвижением ИГ в им. п. (8а–б), либо *это* (7а–б) в позицию подлежащего. Параллелизм (7) и (8) подтверждает, что эксплетив *это* ведет себя как подлежащее в конструкции с подъемом аргумента в главную клаузу. Для диалекта, где допустимы (5е) и (6), речь должна идти не об отсутствии подъема, а о нарушении корреляции между подъемом и приписыванием ТП. Это позволяет выделить два диалекта, отличающиеся настройками микропараметра «подъем аргумента»:

(ii) В А-диалекте, на который опираются существующие модели русской грамматики, подъем подлежащего зависимой клаузы в главную сопровождается заменой ИФ ⇒ ТП. В В'-диалекте подъем подлежащего может реализоваться без замены ИФ ⇒ ТП.

2.6. Сентенциальные аргументы

Корпусные данные свидетельствуют о еще более радикальном отклонении от А-диалекта. В диалекте русской грамматики (далее — В"), который может совпадать или не совпадать с В', ТП приписывается не только *казаться*, *показаться*, но и ненулевыми формами глаголов *быть*, *становиться*, *делаться* в структуре, где они вводят сентенциальный аргумент. Наличие последнего — условие грамматической правильности в В"-диалекте. Примеры типа (9а) регулярно обнаруживаются в XIX–XXI вв., но примеров типа (9б) нет вовсе.

(9) а. Было о ч е в и д н ы м $_{\text{INSTR}}$, [СР что ФНС свободно действовал под крылом и защитой Р. Хасбулатова]. [Вячеслав Костиков. Роман с президентом (1996)].

б. *Было очевидным.

В А-диалекте *быть* и *стать* не являются операторами подъема. Однако дистрибуцию (9а–б) в В"-диалекте трудно объяснить иначе, как подъем всего сентенциального аргумента в позицию подлежащего главной клаузы[4] [Циммерлинг 2018]. Альтернатива — постулировать для (9а-б) скрытое подлежащее ∅, нулевой аналог эксплетива *это*, непривлекательна по двум причинам. Во-первых, в зависимой предикации может быть ненулевое подлежащее, которое не подвергается подъему, ср. ИГ *ФНС* в (9а). Во-вторых, эксплетив *это* выбирается в качестве подлежащего предикатива по иерархии неканонических подлежащих именно тогда, когда сентенциальный аргумент отсутствует, ср. обсуждение в [Zimmerling 2009; 2014; Летучий 2014]. Кроме того, неясно, зачем вводить в описание дополнительную сущность, если ∅ поднимается в главную клаузу лишь при наличии сентенциального аргумента.

Все диалекты русской грамматики подтверждают принцип, в соответствие с которым ТП появляется в силу подлежащного контроля. Позиция подлежащего должна быть заполнена ненулевой формой, а параметрическое варьирование связано с набором разрешенных моделей подъема.

(iii) ТП приписывается именной или адъективной вершине в том случае, когда позиция подлежащего финитного глагола заполнена ненулевым выражением. А-диалект допускает подъем ИГ в им. п. и эксплетива *это* в главную клаузу. В"-диалект также допускает подъем сентенциального аргумента в позицию подлежащего главной клаузы.

[4] SpecTP в традиционной генеративной нотации.

B”-диалект допускает подъем сентенциального аргумента и в случае, когда именной предикат имеет валентность на дат. п. лица, ср. (10а). А-диалект требует ИФ, которая часто интерпретируется как предикатив дативно-предикативной структуры (ДПС), омонимичный согласуемой форме им.-вин. п. ср. р. ед. ч. краткого прилагательного [Поспелов 1955; Циммерлинг 2017].

(10) а. — Мне, — пока еще секретно — сообщили, будто департаменту полиции с т а л о и з в е с т н ы м $_{\text{ADJ.INSTR.SG.N}}$, что вы переслали какое-то письмо отсюда. [Г. А. Гершуни. Из недавнего прошлого (1908)]. — (B”)

б. Департаменту полиции стало и з в е с т н о $_{\text{PRED}}$, что вы переслали письмо. — (A)

в. *Департаменту полиции стало известным.

Характерное для русской лингвистики представление о том, что сентенциальный аргумент всегда занимает позицию подлежащего при матричном предикате, у которого позиция подлежащего не заполнена ИГ в им. п. [Поспелов 1955: 62; Mel’cuk 2014: 183], мотивировано допущением о том, что (10а) и (10б) — варианты той же самой структуры. Это допущение сомнительно, поскольку синтаксис (10а) несовместим с условиями подъема в А-диалекте. Напротив, для (10б) нет доказательств того, что подъем вообще имеет место.

3. Параметр подъема и два диалекта русской грамматики

Построим модель фрагмента русской грамматики, учитывающую варьирование конструкций с разными типами подъема в корпусных данных. Для этого нужно упорядочить микропараметры на шкале рестриктивной силы/стабильности признака:

(iv) $_{\text{Стабильность}}$ [+ подлежащный контроль] >> [+ контроль сентенциальных актантов] … [субституция ИФ ⇒ ТП] >> $_{\text{Варьирование}}$

Запись (iv) означает, что во всех диалектах реализуется общий механизм приписывания ТП, построенный на принципе подлежащного контроля, при этом набор подлежащих, допускающих контроль, и степень обязательности замены ИФ ⇒ ТП различны. Предположим, что диалекты регулируются шкалой (iv). В этом случае носитель А-диалекта не должен одновременно использовать варианты (5е–д) и (10а–б), в то время как в нерестриктивных диалектах B’ и B” это не исключено. Возможность прямого отождествления B’ с B” проверяется путем анализа идиолектов конкретных носителей B’ с B” / массива созданных ими текстов, что представляет собой отдельную проблему[5]. Объединение B’ с B” допустимо рассматривать как единый диалект B, поскольку настройки параметров B’ и B” совместимы.

[5] Для большинства подкорпусов авторских текстов, которые могут быть выделены в НКРЯ, объем представляется недостаточным для окончательного решения этой задачи.

Связка	Схема	Диалект А	Диалект В	
			Диалект В’	Диалект В”
Казалось, показалось	*Казалось… Z-вым/ Z-во, что P*	Прил. в тв. п.	ИФ на *-о*	?
	X-у казалось.. Z-вым/Z-во, что P	Прил. в тв. п.	ИФ на *-о*	?
Было, стало, сделалось	*Было Z-вым/Z-во, что P*	ИФ на *–о*	?	Прил. в тв. п.
	X-у было.. Z-вым/Z-во, что P	ИФ на *–о*	?	Прил. в тв. п.

Рис. 1. Конкуренция исходной формы на ***-о*** и формы тв. п. в диалектах русской грамматики

3.1. Статус исходной формы на *–о*

Для А-диалекта, где подъем сентенциального аргумента не допускается, можно отождествить ИФ в предложениях типа *(X-у) было очевидно, что P* с несогласуемым предикативом и считать, что от предикативов на *–о,* в отличие от кратких прилагательных на *–о*, формы ТП образованы быть не могут. Данная точка зрения была впервые высказана в [Поспелов 1955], см. также [Циммерлинг 2018]. Оговорка Н. С. Поспелова о том, что пары предложений *Мне было грустно* vs *Мне было грустно, что P* якобы реализуют омонимы предикатив : прилагательное, не нужна. Отсутствие преобразования ИФ ⇒ ТП при предикативах ДПС типа *известно* в А-диалекте подкрепляет гипотезу о том, что сентенциальный аргумент не выбирается в А-диалекте в качестве подлежащего предикатива ДПС при внешне выраженном дат. п. лица [Циммерлинг, Трубицина 2015].

Для В”-диалекта, где допустимо варьирование *(X-у) известно/известным, что P,* сохраняется двойственность описания. ИФ можно рассматривать либо как предикатив — в этом случае имеет место конкуренция несогласуемого неконтролируемого предикатива с контролируемым прилагательным, либо как прилагательное — в этом случае варьирование *было известно/известным, что P* свидетельствует о факультативности замены ИФ ⇒ ТП после БЫЛО, СТАЛО, аналогично тому, как В’-диалект допускает оба варианта *казалось неумным/неумно* после КАЗАТЬСЯ.

3.2. Исходная форма на *–о* при КАЗАТЬСЯ

Конструкция с ИФ после КАЗАТЬСЯ возможна только в В-диалекте, ср. пример (6), повторенный ниже как (11а). Значение (11а) и структуры с финитным придаточным (11б), возможной как в А-, так и В-диалекте, идентично.

(11) а. Это показалось мне сентиментально, приторно и неумно, (А. П. Чехов) — В-диалект.
б. Мне показалось, [СР что это сентиментально, приторно и неумно].

(12) *Мне было сентиментально, приторно и неумно.

Следует заключить, что в (11а) реализуется особый случай подъема, без замены ИФ ⇒ ТП. Критерий для проверки дает отсутствие контрпримеров типа (12) в корпусе — слова *сентиментально* и *неумно* не являются предикативами ДПС, а (11а) передает значение ‘X счел, что это сентиментально и неумно’, а не значение ‘X пришел в сентиментальное настроение и повел себя неумно’.

4. А- и В-диалекты в НКРЯ

Для оценки количественного соотношения А- и В-диалектов в НКРЯ был проведен эксперимент. На первом этапе на базе 7 пар вида ИФ vs ТП проверялась конструкция с глаголами *быть*, *стать*, *делаться* и подъемом *что Р*-аргумента, возможная в В-диалекте. На втором этапе те же 7 пар стимулов проверялись в контексте *что Р*-аргумента и глаголов *казаться*, *показаться*. На третьем этапе проверялись комбинации *быть* + сентенциальный аргумент + ТП и *казаться* + сентенциальный аргумент + ИФ со стимулами разной семантики.

4.1. Подъем сентенциального аргумента в В-диалекте

Мы протестировали соотношение ТП vs ИФ в 7 парах: *очевидным/очевидно*, *явным/явно*, *ясным/ясно*, *понятным/понятно*, *известным/известно*, *странным/странно*, *сомнительным/сомнительно*. Учитывалось 5 форм связок: *было*, *стало*, *становится*, *сделалось*, *делается*. Отдельно подсчитывались употребления с внешне выраженным дат. п. лица. Статистика отражены на рис. 2, темным выделены показатели конструкции ДПС. Выбор формы ТП однозначно свидетельствует о реализации В-диалекта, в то время как выбор ИФ, при сделанном выше допущении о том, что В-диалект включает в себя варианты, возможные в рестриктивном А-диалекте, указывает на А ∩ В.

Вероятность выбора структуры с ТП и подъемом сентенциального аргумента выше всего при неупотребительных связках *сделалось, делается* (54,5 %). Для *стало, становится* вероятность выбора структуры с ТП в предложении без дат. п. лица составляет 12,7 %, а для *было* — 1,2 %. Наличие внешне выраженного дат. п. лица — фактор, понижающий вероятность подъема сентенциального аргумента (в среднем — 1,9 %), при связке *было* и дат. п. лица такая структура не встретилась вовсе. Тем самым, корпусные данные подтверждают выделенное положение конструкции ДПС в русской грамматике.

Эксперимент показал, что конструкция с ТП и сентенциальным аргументом при ненулевых формах связок *быть, стать, становиться* довольно распространена, хотя конкурирующая с ней конструкция с ИФ намного более частотна. Величина выгрузки примеров с ТП для каждой пары стимулов ниже соответствующего показателя для ИФ, но есть комбинации, где вероятность

появления ТП существенно выше среднего значения. Сюда относятся комбинации *становится очевидным, что P* — 88 примеров при 38 примерах *становится очевидно, что P*; *становится понятным, что P* — 42 примера при 46 примерах *становится понятно, что P*; *стало очевидным, что P* — 100 примеров при 163 примерах *стало очевидно, что P*.

Связка	Схема	Прилагательное в тв. п. <B-диалект>		Предикатив/ИФ на *–о* <A ∩ B>	
было	*было.. Z-вым/Z-во, что P*	32		2529	
	X-у было.. Z-вым/Z-во, что P	0		995	
стало, становится	*стало.. Z-вым/Z-во, что P*	406		2788	
	X-у стало Z-вым/Z-во, что P	26		636	
сделалось, делается	*сделалось Z-вым/Z-во, что P*	54		45	
	X-у сделалось Z-вым/Z-во, что P	6		16	
		Всего: 524	с дат. п.: 32	Всего: 7009	ДПС: 1647

Рис. 2. Конкуренция предикативов и прилагательных в контексте *было/стало Z-вым/Z-во, что P*

4.2. Исходная форма на *–о* с КАЗАТЬСЯ в B-диалекте

Конструкции с КАЗАТЬСЯ реализуются в НКРЯ реже, что показала выборка с теми же 7 парами стимулов.

Связка	Схема	Прилагательное в ТП <A-диалект>		Предикатив/ИФ на *–о* <A ∩ B>	
казаться, показаться	*казалось.. Z-вым/Z-во, что P*	78		7	
	X-у казалось... Z-вым/Z-во, что P	213		9	
		Всего: 291	с дат. п.: 213	Всего: 16	ДПС: 9

Рис. 3. Предикативы и прилагательные в контексте *казалось Z-вым/Z-во, что P*

В XIX в. 6 раз встретилась комбинация *кажется ясно, что P*:

(13) Но из первой моей статьи к а ж е т с я я с н о , ч т о вопрос идет не о достоверности летописи вообще, а только о некоторых начальных её известиях, [Д. И. Иловайский. Начало Руси (1876)]

Зафиксировано 4 примера *X-у кажется ясно, что P*, из них 2 относятся к 1-й половине XX в., ср. (14).

(14) Он действительно всегда забывал давать ей деньги: так е м у к а з а л о с ь я с н о , ч т о деньги у них общие. [М. А. Алданов. Истоки. Части 9–17 (1942–1946)]

4.3. Дативная конструкция с КАЗАТЬСЯ

Предикативы типа *ясно* имеют валентность на дат. п. лица, поэтому у (12) есть две интерпретации:

(14') Х-у б ы л о я с н о , что деньги у них общие. (ДПС)

(14'') Х с ч и т а л я с н ы м , что деньги у них общие. (структура с подъемом)

Для верификации гипотезы о том, что сочетание *Х-у кажется Z-во* может отражать подъем сентенциального аргумента, надо взять предикатив, лишенный стандартной валентности на дат. п. лица. С расширенным набором стимулов проверка дала ок. 30 примеров в 1813–2005 гг. Ср. два примера, разделенные почти 150 годами:

(15) Не знаю как у других, но я его вижу на своем лице и м н е э т о к а ж е т с я н е к р а с и в о , как маска какая-то. [Красота, здоровье, отдых: Косметика и парфюм (форум) (2004)]

(16) — Если в а м к а ж е т с я н е к р а с и в о э т о , то не глядите и отворачивайтесь. [А. Ф. Писемский. Люди сороковых годов (1869)]

Конструкция с подъемом и ИФ при КАЗАТЬСЯ засвидетельствована также с *логично, прикольно, грязно, неразумно, невежливо, немыслимо, губительно, непедагогично, сумрачно, недостойно, придирчиво, сентиментально, неумно, безрассудно, необъяснимо, бесплодно, неучтиво, благоразумно, мудрено*.

4.4. Маргинальные конструкции с КАЗАТЬСЯ

Помимо регулярных случаев употребления ИФ краткого прилагательного при *казаться*, НКРЯ фиксирует спорадические примеры других употреблений форм им. п., где подстановка ТП невозможна или затруднена. Сюда относятся употребления полной формы прилагательного, ср. *Мне к а ж е т с я и н о е* [К. Н. Леонтьев, 1888] ‘Х полагает, что дело обстоит иначе’, *отвращаться от того, что кажется ей худое* [Я. П. Козельский, 1768] ‘отказываться от того, что кажется ей плохим’ и употребления глагола *казаться* в значениях ‘померещиться, почудиться’ — *Мне кажется страшное* [А. П. Платонов, 1933–35] и ‘нравиться’ — *Ну что, братец, как вам кажется здешнее местоположение?* [Д. Т. Ленский, 1833]. В этих случаях *казаться* не приписывает ТП в структуре, которая может быть интерпретирована как подъем нульместной предикации с опущенным инфинитивом *быть*.

В XVIII в. инфинитив в структуре с подъемом мог быть выражен эксплицитно:

(17) Время летнее к а ж е т с я б ы т ь у г о д н е й ш е е к п о с е щ е н и ю , нежели зимнее. [[Петр I]. Регламент или устав духовной коллегии (1721)]

Подъем ИГ *время летнее* в позицию подлежащего в (17) несовместим с современной русской грамматикой. Напротив, подъем в высказываниях типа (18), с инфинитивом при предикативе ДПС, у которого позиция подлежащего не заполняется ИГ в им. п., например, *скучно* в (18), соответствует В-диалекту.

(18) на сии предварения, госпоже Руссо
п о к а з а л о с ь с к у ч н о о ж и д а т ь кончины своего супруга
[Д. И. Фонвизин. Письма П. И. Панину (1777–1778)]
'Х-у показалось, что ожидать р — скучно'.

5. Выводы и перспективы

Была построена модель фрагмента русской грамматики, предсказывающая варьирование конструкций с подъемом аргумента и распределение вариантов с ТП и ИФ в корпусных данных. Русский язык XVIII-XXI вв. в данной области характеризуется двумя тенденциями — a) ограничением числа допустимых конструкций с подъемом аргумента, b) установлением дополнительной дистрибуции прилагательных в ТП и несогласуемых предикативов, что достигается в А-диалекте. Выделенное положение конструкции ДПС, выражающей значение внутреннего состояния — эмпирическая основа теорий Л. В. Щербы и Н. С. Поспелова, которые представили предикативы ДПС как класс слов, отличных от прилагательных. В пользу анализа Щербы — Поспелова свидетельствует тот факт, что в А-диалекте подъем сентенциального аргумента предикатива ДПС в позицию подлежащего затруднен или невозможен. В-диалект сохраняет возможность подъема сентенциального аргумента при предикативах всех типов.

Литература

1. *Гловинская 1998* — Гловинская М. Я. Активные процессы в грамматике (на материале инноваций и массовых языковых ошибок) // Русский язык конца XX столетия (1985–1995). М.: Языки славянской культуры, 1996. С. 217–304.
2. *Красильникова 1982* — Красильникова Е. В. О соотношении языковых уровней в системе русской разговорной речи // Проблемы структурной лингвистики 1980. М., 1982. 37–49.
3. *Лаптева 1976* — Лаптева О. П. Русский разговорный синтаксис. М.: Наука, 1976.
4. *Летучий 2014* — Летучий А. Б. Синтаксические свойства сентенциальных актантов при предикативах // Вестник МГГУ им. М. А. Шолохова. Сер. Филологические науки. 2014, № 1. С. 62–84.
5. *Лютикова, Циммерлинг, Коношенко 2016* — Лютикова Е. А., Циммерлинг А. В., Коношенко М. Б. Языковое разнообразие в зеркале параметрической грамматики // Е. А. Лютикова, А. В. Циммерлинг, М. Б. Коношенко (ред.). Типология морфосинтаксических параметров, Вып. 3. Материалы международной конференции «Типология морфосинтаксических параметров 2016. М.: МПГУ, 2016. С. 5–15.
6. *Попов 1881* — Попов А. В. Синтаксические исследования. Воронеж, 1881.

7. *Ожегов 1964* — Ожегов С. И. Словарь русского языка. Изд-е 6-е, стереотипное. М.: Советская энциклопедия, 1964.
8. *Поспелов 1955* — Поспелов Н. С. В защиту категории состояния // Вопросы языкознания. 1955. № 2. С. 55–65.
9. *Циммерлинг 2012* — Циммерлинг А. В. Неканонические подлежащие в русском языке // От значения к форме, от формы к значению: Сборник статей в честь 80-летия Александра Владимировича Бондарко. М.: Языки славянской культуры, 2012. С. 568–590.
10. *Циммерлинг 2017* — Циммерлинг А. В. Русские предикативы в зеркале эксперимента и корпусной грамматики // Компьютерная лингвистика и интеллектуальные технологии, вып. 15 (23). М.: РГГУ, 2017. Т2. С. 466–482.
11. *Циммерлинг 2018* — Циммерлинг А. В. Предикативы и предикаты состояния в русском языке // Slavistična revija, 2018, № 1, С. 45–64.
12. *Циммерлинг, Трубицина 2015* — Циммерлинг А. В., Трубицина М. В. Дативные и сентенциальные подлежащие в русском языке: от внутренних состояний к общим суждениям // Rhema. Рема. 2015, №. 4. С. 71–104.
13. *Шведова 1982* — Шведова Н. Ю. (отв. ред.). Русская грамматика в 2 т. М.: Наука, 1982.
14. *Щерба 2008* — Щерба Л. В. Языковая система и речевая деятельность. 4-е Изд. М.: ЛКИ, 2008.
15. *Wilson, Henry 1998* — Wilson J. and Henry A. (1998). Parameter setting within a socially realistic linguistics // Language in Society, Vol. 27, No. 1, 1–21.
16. *Henry 1998* — Henry A. (1998). Dialect variation, optionality and the learnability guarantee // Linguistica Atlantica, 20 (1998), 51–71.
17. *Harris 1996* — Harris J. (1996). Syntactic variation and dialect divergence // Singh Rajendra (ed.), Towards a critical sociolinguistics. Amsterdam: Benjamins, 31–59.
18. *Bailyn 2012* — Bailyn J. Syntax of Russian. Cambridge: Cambridge University Press, 2012.
19. *Comrie, Stone, Polinsky 1996* — Comrie B., Stone G. & M. Polinsky. Russian Language in the 20th century. Oxford: Clarendon Press, 1996.
20. *Matushansky 2008* — Matushansky O. A Case Study of Predication // F. Marušič and R. Žaucer, eds., Studies in Formal Slavic Linguistics. Contributions from Formal Description of Slavic Languages 6.5. Frankfurt am Main: Peter Lang, 213–239.
21. *Mel'cuk 2014* — Mel'cuk I. Syntactic Subject: Syntactic Relations, Once Again // В. А. Плунгян, М. А. Даниэль, Е. А. Лютикова, С. Г. Татевосов, О. В. Федорова (ред.). Язык. Константы. Переменные. Памяти А. Е. Кибрика. Санкт-Петербург: Алетейя, 2014. С. 169–216.
22. *Zimmerling 2009* — Zimmerling A. Dative subjects and semi-expletive pronouns in Russian // Studies in Formal Slavic Phonology, Morphology, Syntax, Semantics and Discourse Structure [Linguistik International, Band 21] /G. Zybatow, U. Junghanns, D. Lenertova, P. Biskup (eds.). Peter Lang, 2009, 253–265.
23. *Zimmerling 2014* — Zimmerling A. Sententional arguments and event structure // Компьютерная лингвистика и интеллектуальные технологии, выпуск 13 (20). По материалам ежегодной международной конференции «Диалог» (2014). М: РГГУ, 2014. С. 710–727.

РАЗРАБОТКА МОДЕЛИ КОММУНИКАТИВНОГО ПОВЕДЕНИЯ РОБОТА Ф-2 НА ОСНОВЕ МУЛЬТИМОДАЛЬНОГО КОРПУСА «REC»[1]

Зинина А. А. (zinina_aa@nrcki.ru),
Аринкин Н. А. (arinkin_na@nrcki.ru),
Зайдельман Л. Я. (zaydelman_ly@nrcki.ru),
Котов А. А. (kotov_aa@nrcki.ru)
НИЦ «Курчатовский институт», Москва

В статье описывается разрабатываемая архитектура для моделирования естественного коммуникативного поведения на роботе Ф-2. Важной частью нашей работы является корпусное исследование коммуникативного поведения человека и последующий перенос такого поведения на робота. Основываясь на мультимодальном корпусе REC, мы описываем особенности естественной коммуникации, а также разрабатываем архитектуру, которая учитывает такие особенности. В данной архитектуре робот может по-разному выражать какую-либо коммуникативную функцию, используя один или несколько исполнительных органов: например, демонстрировать *апелляцию* с помощью мимики, движений головы или жестов рук. Разработанная архитектура также позволяет гибко комбинировать жесты с разными коммуникативными функциями. Архитектура позволяет с помощью режимов split, join и single комбинировать теги из разных BML-пакетов, а также синхронизировать теги внутри одного пакета BML. Перечисленные особенности являются ключевыми для формирования правдоподобного поведения робота Ф-2 и необходимы для повышения эффективности коммуникации между роботом и пользователем.

Ключевые слова: мультимодальная коммуникация, коммуникативные функции, архитектура робота-компаньона

[1] Разработка системы обработки речи для робота поддержана грантом РФФИ 16-29-09601 «Система автоматического выявления эмоциональных и экстремистских суждений в текстах на естественном языке».

DEVELOPMENT OF COMMUNICATIVE BEHAVIOR MODEL FOR F-2 ROBOT BASING ON "REC" MULTIMODAL CORPORA

Zinina A. A. (zinina_aa@nrcki.ru),
Arinkin N. A. (arinkin_na@nrcki.ru),
Zaydelman L. Ya. (zaydelman_ly@nrcki.ru),
Kotov A. A. (kotov_aa@nrcki.ru)
Kurhcatov Institute, Moscow, Russia

The article describes the developed architecture for modeling natural communicative behavior on the F-2 robot. The important part of our work is the study of human communicative behavior and the transfer of this behavior to the robot. For this purpose we are developing the Russian Emotional Corpus (REC) where video recordings of natural emotional dialogues are collected. We explore the features of natural communication, and also develop an architecture that takes into account these features. For example, using the architecture presented in the article a robot can express any communicative function, using one or more executive organs: for example, to express an *appeal* with facial expressions, head movements or gestures. The developed architecture also allows us to flexibly combine gestures with different communicative functions. The architecture allows us to use "split", "join" and "single" modes to combine tags from different BML-packages, and also to synchronize tags in a single BML-package. These features are important for modeling of human-like behavior for the robot F-2, and are necessary to improve the communication between a robot and a user.

Keywords: multimodal communication, communicative functions, architecture of the companion robot

1. Введение

В современной робототехнике проводится большое количество исследований, посвященных взаимодействию между роботом и пользователем. Исследователи пытаются оценивать эффективность такого взаимодействия, а также повышать удовлетворенность пользователя от общения с роботом. Чтобы взаимодействие человека и робота происходило естественным образом, роботы должны использовать выразительные средства, близкие пользователю.

С одной стороны, неотъемлемой частью социального взаимодействия между роботом и человеком является эмоциональная модель робота. Например, в работе [Kirbya, Forlizzib и др, 2010] авторы сфокусированы на моделировании долгосрочных аффективных состояний робота, они показывают эффективность этой модели в различных социальных ситуациях. [Lee, Ahn и др., 2009] генерируют эмоциональную матрицу, которая представляет возможности выражения модульного поведения на основе эмоциональных следов. [Velásquez, 1998]

разрабатывает архитектуру, где первичные эмоции выступают в качестве строительных блоков для эмоциональных воспоминаний, которые играют существенную роль в процессе принятия решений и выборе действий.

С другой стороны, [Breazeal, Scassellati, 2002] утверждают, что имитация роботом поведения живого человека является важным инструментом при установлении контакта между роботом и собеседником. Результаты эксперимента [Leite, Pereira и др., 2008] показали, что эмоциональное поведение повышает эффективность взаимодействия между физическим роботом и пользователем. Исследователи [Beuter, Spexard и др., 2008] отмечают существенную значимость гибкого комбинирования жестов и речи при «интуитивном» человеко-машинном взаимодействии. Именно последнее важно для формирования эмоционального контакта с роботом, а также для повышения удовлетворенности пользователей при таком взаимодействии. Поэтому моделирование правдоподобного коммуникативного поведения робота, включая мимику и жесты, является важной и перспективной задачей. Более того, богатая коммуникативная модель робота является важным конкурентным преимуществом робота среди существующих аналогов. В статье будут описаны особенности построения и практической реализации данной модели на роботе Ф-2.

В лаборатории нейрокогнитивных технологий Курчатовского комплекса НБИКС-технологий мы разрабатываем проект робота Ф-2, который включает синтаксический парсер — для автоматического выделения из текста существенных семантических компонентов [Kotov, Zinina, Filatov 2015], а также систему управления коммуникативным поведением — для жестовых и мимических реакций робота на входящие тексты (**Рис. 1**). Мы также разрабатываем гибкую эмоциональную модель робота, которая основана на работе 13 негативных сценариев (Опасность, Обман, Неадекватность антагониста, Эмоциональность антагониста и т. д.) и 23 позитивных (Забота, Контроль над ситуацией, Служение и т. д.) [Kotov, 2003, 2012].

Представленная система носит модульный характер, она включает в себя обработку поступающего на вход звукового или текстового сообщения. Полученное сообщение передается в модуль морфологической обработки, затем — в модуль синтаксической обработки. Результаты работы модуля — синтаксические деревья — передаются в последующий компонент семантического анализа. Для каждой клаузы в синтаксическом дереве строится семантическое представление — множество признаков, распределённых по семантическим валентностям: *предикат*, *агенс*, *пациенс* и т. д. Каждый сценарий включает аналогичные семантические структуры — множества признаков, распределённых по валентностям. Для каждой пары вида <семантическое представление, сценарий> вычисляется мера близости, она зависит от числа совпавших семантических признаков в тождественных валентностях [Kotov, Zinina, Filatov 2015]. На основе меры близости и чувствительности сценариев вычисляется активизация каждого сценария. Наиболее активировавшийся сценарий выражается через коммуникативные функции, передавая на робота жесты, элементы мимики и речь в формате BML (Behavior Markup Language) [Kopp, Krenn и др., 2006; Vilhjálmsson, Cantelmo и др. 2007]. Подробнее система управления роботом Ф-2 описана в [Kotov, Arinkin и др., in press].

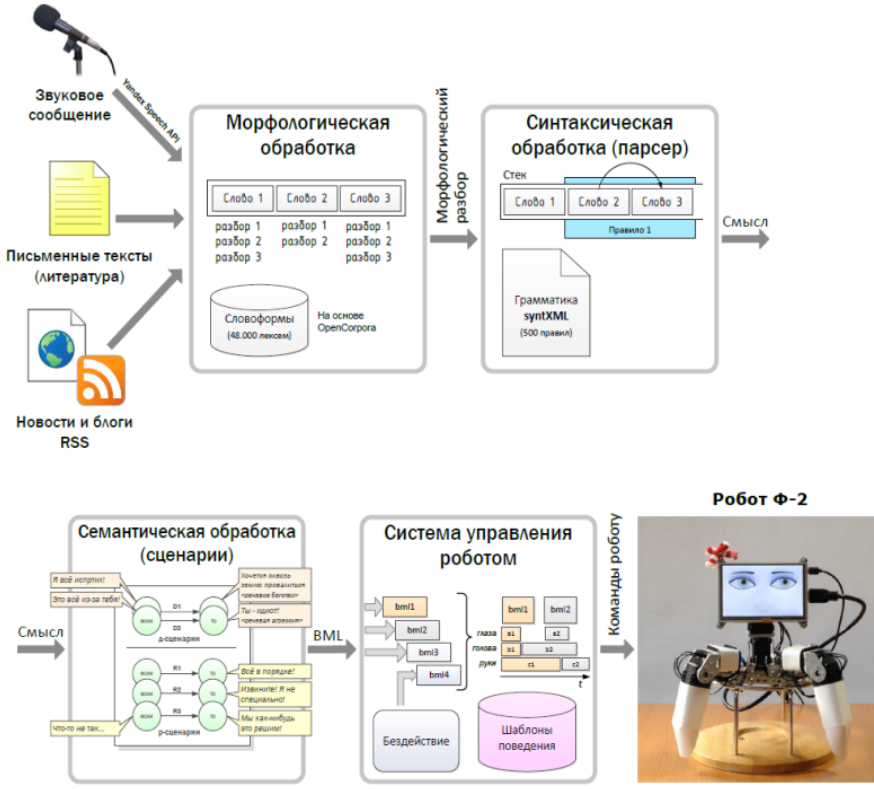


Рис. 1. Общая схема обработки входящего текста и синтеза поведенческих реакций роботом Ф-2

Важной частью нашей работы является корпусное исследование коммуникативного поведения человека и последующий перенос такого поведения на робота. Можно сказать, что в рамках работы мы решаем как теоретические (описание и классификация коммуникативного поведения человека), так и прикладные (перенос такого поведения на робота) задачи. Важно отметить, что робот Ф-2 обладает достаточно простой архитектурой тела — мы не стремимся к полной имитации роботом тела человека. Однако именно естественные мимика и жесты вкупе с гибким речевым поведением робота обеспечивают эффект, аналогичный эмоциональной симпатии к мультипликационным персонажам, когда симпатию может вызывать даже с виду «странный» герой.

Коммуникативное поведение человека мы исследуем на основе мультимодального корпуса REC (Russian Emotional Corpus), который содержит размеченные в программе ELAN [Brugman, Russel, 2004] видеозаписи эмоциональных диалогов на университетских экзаменах (295 фрагментов), в муниципальной службе одного окна (510 фрагментов), а также диалогов с информантами, которые занимаются каким-либо видом искусства, например, хореографией или

рисованием (10 фрагментов). Разработка корпуса осуществляется с 2008 года [Kotov, Budyanskaya, 2012]. В корпусе вручную размечаются речевые высказывания участников диалога. Для информанта (студента, клиента, респондента) размечаются движения глаз, губ и рук. Мимикой глаз считаются коммуникативно значимые движения глазами верх, вбок, расширение глаз, частые моргания и др. К движениям губ относятся улыбки, облизывания, прикусывания и др. Разметка движений рук выполняется на 4-х слоях: это способ, активный и пассивный органы, а также траектория. Здесь отмечены всевозможные почесывания, поглаживания, манипуляции, иконические знаки и др.

Отдельный уровень разметки задает коммуникативную функцию мимического движения или жеста, если эта функция может быть определенно установлена [Kotov, Zinina, 2015a; Kotov, Zinina, 2015b]. Коммуникативные функции приписывают новые параметры аннотациям из базовой разметки корпуса, а также классифицируют движения головы и тела человека, поскольку базовая разметка у этих элементов в корпусе отсутствует. На основании результатов контент-анализа базовых элементов разметки мы выделяем 35 тегов функциональной разметки. Например, отдельно отмечаем *понимание-согласие-одобрение*, *отрицание-несогласие-возражение*, *апелляцию*, *побуждение*, *ожидание-обратной-связи*, *остановку-адресата*, *отсутствие-невозможность* и другие. При разметке видеозаписей мы специально отслеживаем согласованность экспертов, то есть отдельную видеозапись просматривают не менее двух человек, после чего обсуждаются некоторые спорные или неоднозначные моменты.

Анализ коммуникативных функций позволяет выделить типичные элементы поведения, характерные для той или иной функции, или инвариант поведения — эти элементы зарисовываются в 3D-редакторе Blender (**Рис. 2**) и сохраняются в базу данных MySQL.

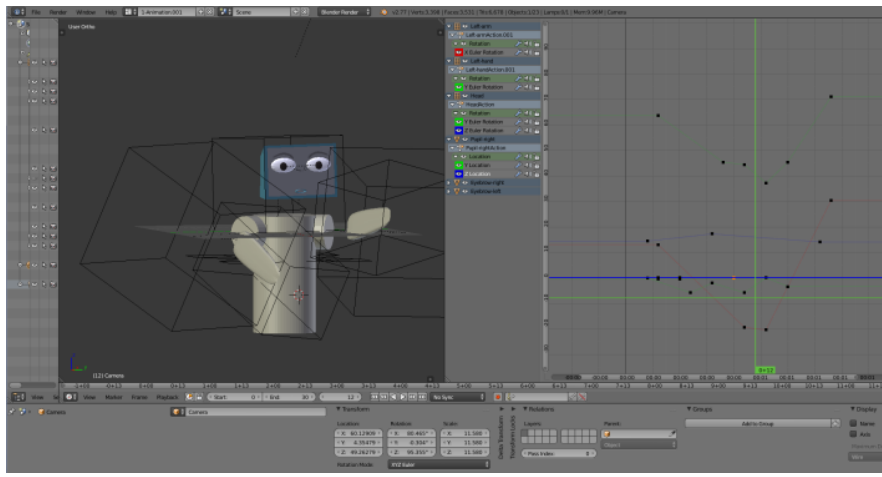


Рис. 2. Разработка жеста для робота Ф-2 в 3D-редакторе Blender

Далее — жест может быть извлечен из базы с помощью скриптов на языке BML — (**Рис. 3**). Отдельные теги внутри пакета BML обозначают движения каждого исполнительного органа робота (движения рук, головы, элементов лица), а также речь. В работах [Kopp, Krenn и др., 2006; Vilhjálmsson, Cantelmo и др., 2007] поведение кодируется с помощью длинного скрипта BML, однако мы создаем отдельный BML-пакет для каждой коммуникативной реакции, что позволяет гибко комбинировать различные элементы поведения на роботе с целью моделирования разнообразного и правдоподобного поведения.

```
<bml id="1" syncmode="single">
    <head id="4" lexeme="appeal3" start="1:start"/>
    <pupils id="3" lexeme="appeal3" start="1:start"/>
    <hands id="2" lexeme="appeal3" start="1:start"/>
    <speech id="1">
            Неужели нужно так нервничать?
    </speech>
</bml>
```

Рис. 3. Скрипт на языке BML

По результатам работы мы формируем словарь жестов с анимированными иллюстрациями исходного движения из корпуса и того же движения, разработанного для робота. Подобный словарь существенно упрощает работу по моделированию коммуникативного поведения на роботе.

2. Работа с мультимодальным корпусом

Анализ коммуникативного поведения человека необходим при моделировании эмоционального поведения робота. В корпусе REC представлены реальные эмоциональные ситуации, и воспроизводя коммуникативное поведение информантов, робот может повысить правдоподобность своих действий в диалоге. Существенной особенностью поведения людей в корпусе является то, что они демонстрируют вариативные и множественные коммуникативные реакции — человек может отвечать сразу несколькими часто противоречивыми реакциями на входящее высказывание. Поэтому наша задача при разработке модели поведения робота состоит в том, чтобы обеспечить разнообразие в выражении отдельной коммуникативной функции, а также позволить роботу одновременно выполнять несколько различных коммуникативных действий.

2.1. Инварианты коммуникативных функций

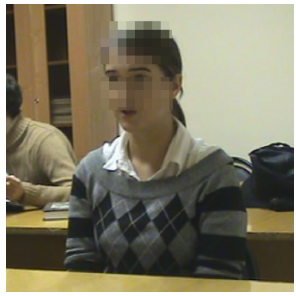
В корпусе большинство коммуникативных функций не привязаны к единственному способу выражения. Наоборот, существуют определенные тенденции или ограничения на выражение коммуникативных функций. Например,

такая функция как *апелляция* в 15,8% случаев выражается с помощью мимики, в 46,1% — движениями головы, в 24,5% — жестами рук, и в 13,6% — движениями тела, а *компенсация-закрытие* с помощью мимики выражается в 31,6% случаев, движениями головы — лишь в 1,9%, жестами рук — 62,7%, движениями тела — 3,8% [Kotov, Zinina, 2015b].

Вместе с тем, анализ разметки позволяет выделить спектр типичных представителей отдельной коммуникативной функции, например, кивок головой или движение вниз кистью руки для выражения *понимания-согласия-одобрения*. Такие типичные представители определенной коммуникативной функции включаются в цикл порождения движений роботом.

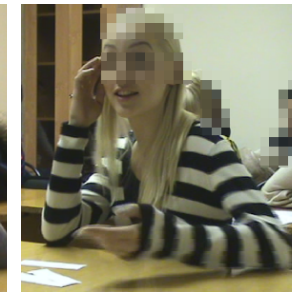
2.2. Комплексные и редуцированные жесты

Определенная коммуникативная функция может быть выражена не только разными исполнительными органами, но и сложной комбинацией из нескольких исполнительных органов. Например, *апелляцию* можно выразить только с помощью мимики — поднятием бровей (**Рис. 4а**), только с помощью движения головы — наклоном головы вперед или кивком вниз, или только с помощью движения руки — протягиванием руки к адресату, а также через сложную комбинацию жестов — например, через движения головы и мимики (**Рис. 4б**), руки и мимики или головы, а также, что реже, с помощью комплексного движения головы, мимики и руки (**Рис. 4в**).



(а) выражение *апелляции* через мимику (поднимает брови), 20081225-zhurn-b6-m (10:01.020)

(б) выражение *апелляции* через мимику и движение головы (поднимает брови и наклоняется вперед), 20081219-zhurn-a02 (02:38.040)



(в) выражение *апелляции* через комплексное движение с помощью головы, мимики и рук (поднимает брови, наклоняется вперед, поднимает подбородок вверх, направляет руку к собеседнику), 20081225-fipp-a17-m (02:58.313)

Рис. 4. Выражение апелляции с помощью мимики; мимики и головы; мимики, головы и руки

2.3. Пересечение коммуникативных функций

В реальном коммуникативном поведении можно наблюдать примеры, где информант одновременно выражает несколько различных коммуникативных функций — в этом случае мы говорим об их наложении. Например, информант в случае 20081225-fipp-a02 (01:16.834) одновременно выражает *я-размышление* («задумывается») с помощью поворота головы и прищуривания глаз, а также демонстрирует функцию *остановка-адресата* движением руки (**Рис. 5**).

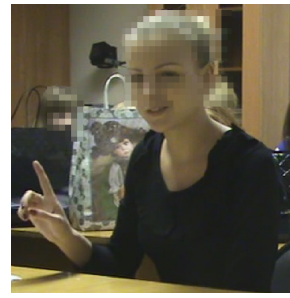


Рис. 5. Совмещение коммуникативных функций:
я-размышление («задумчивость») выражается головой и глазами,
остановка-адресата выражается движением руки, 20081225-fipp-a02 (01:16.834)

Исходя из диаграммы (**Рис. 6**) можно сделать вывод, что такая функция как *яразмышление*, как правило, пересекается с такими функциями как: *речь-эмфаза, дистанцирование, отрицание-несогласие-возражение, отсутствие-невозможность, стимулирование*. Причем *я-размышление* в таких примерах выражается преимущественно с помощь мимических паттернов, а другие коммуникативные функции — с помощью движений рук, головы и тела.

Таким образом, при разработке системы для синтеза правдоподобного поведения робота Ф-2 необходимо учитывать перечисленные особенности реальной коммуникации.

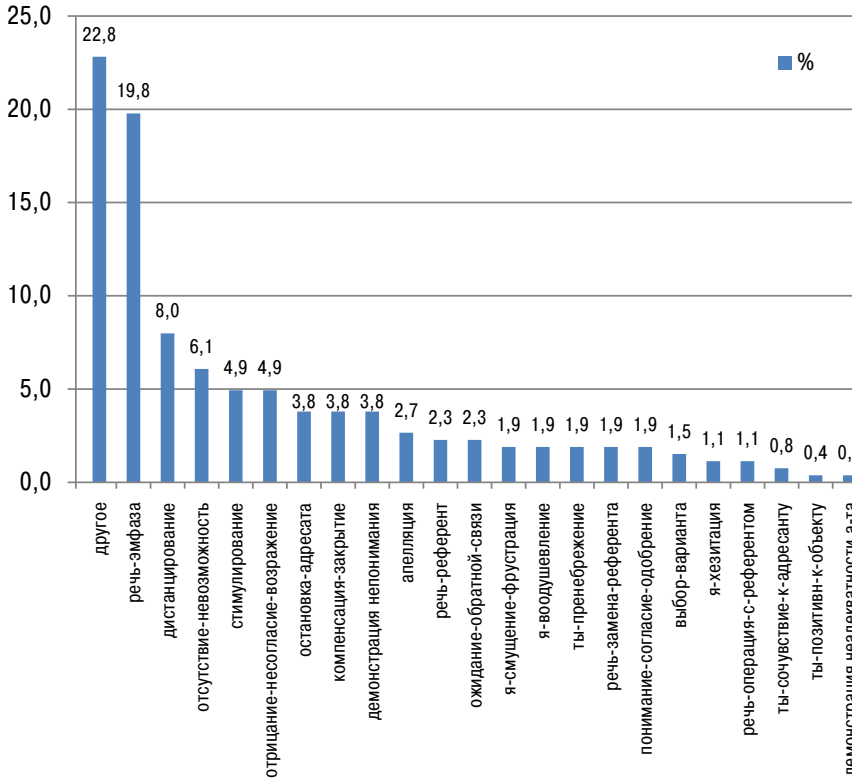


Рис. 6. Диаграмма пересечения коммуникативной функции *я-размышление* с другими коммуникативными функциями

3. Особенности архитектуры робота Ф-2

В рамках работы над проектом робота Ф-2 мы разрабатываем такую архитектуру, которая в зависимости от активации определенного сценария продуцирует коммуникативные реакции, вызывающие BML-пакеты. В зависимости от активации определенного сценария робот может порождать различный ответ — использовать комбинацию разных исполнительных органов и речь. Выражение коммуникативной функции снижает активизацию соответствующего сценария, однако если эта активизация все еще находится выше порогового значения, робот может дополнительно выразить коммуникативную функцию с помощью других элементов поведения. Для каждой коммуникативной функции в базе хранится несколько BML-пакетов, которые кодируют различные способы ее выражения. На сегодняшний момент в базе сохранено 577 BML-пакетов. BML-пакет включает набор тегов, связанных с коммуникативной функцией (**Рис. 3**). С помощью тега описываются исполнительные

органы (глаза, голова, речь, левая и правая рука), на которых будет воспроизведен жест, соответствующий определенной коммуникативной функции.

Жесты робота представляют собой последовательность временных меток, к каждой из который приписаны определенные инструкции. При активации тега робот последовательно обрабатывает метки и связанные с ними инструкции передаются для исполнения на приводы робота (для движения рук и головы), на экран (для мимики) и на аудиосистему (для синтеза речи). BML-пакеты жестов могут использовать только часть тегов, чтобы жесты из разных коммуникативных функций могли исполняться одновременно. Таким образом, при комбинировании коммуникативных функций у нарисованного жеста могут вызываться только некоторые теги BML. Например, у жеста с функцией *я-размышление* активируются временные метки, связанные с движением головы и мимикой, а у жеста с функцией *остановка-адресата* — с движением руки. Робот одновременно «задумывается», смотрит вбок — и выполняет отрицательный жест рукой (**Рис. 7**). По завершении обработки сценария его активизация уменьшается и его BML-пакеты удаляется из очереди. Таким образом, основанная на тегах архитектура позволяет гибко комбинировать жесты с различными коммуникативными функциями, что, в свою очередь, значительно обогащает поведение робота.

Рис. 7. Совмещение коммуникативных функций на роботе Ф-2:
я-размышление (выражается головой и глазами)
и *остановка-адресата* (выражается жестом руки)

3.1. Очередь BML-пакетов

BML-пакеты, порожденные различными коммуникативными функциями, направляются в очередь и исполняются по мере того, как на роботе освобождаются требуемые исполнительные органы. Архитектура робота позволяет по-разному формировать очередь из BML-пакетов за счет того, что

у каждого BML-пакета присутствуют параметр сочетания с предшествующими пакетами — syncmode. Установка значений данного параметра позволяет по-разному группировать BML-пакеты. Можно выделить три режима, задаваемых при помощи syncmode — split, join и single.

Режим single гарантирует, что выполнение пакета BML будет запущено, только если все исполнительные органы робота свободны. Это обеспечивает выполнение одного BML-пакета без пересечения по времени с другими BML-пакетами (**Рис. 8**).

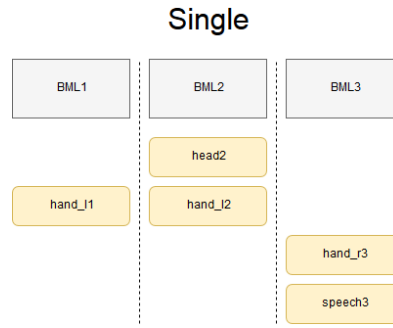


Рис. 8. Режим single: последовательно будут выполнены теги пакетов BML1, BML2 и BML3

Режим split позволяет запустить обработку и выполнение тега из следующих в очереди BML-пакетов, если требуемый тегу исполнительный орган освободился. То есть пакеты не будут ждать в очереди, пока закончится выполнение тегов предыдущего BML, а будут конкурировать за свободные исполнительные элементы. Таким образом, выполнение тегов одного BML-пакета может быть дополнено тегами других пакетов, что дает разнообразие в поведении робота с помощью комбинирования жестовых движений. На **Рисунке 9** изображена последовательность выполнения тегов BML в очереди по порядку, а справа — очередь робота, в которой каждому из трех пакетов установлен режим split.

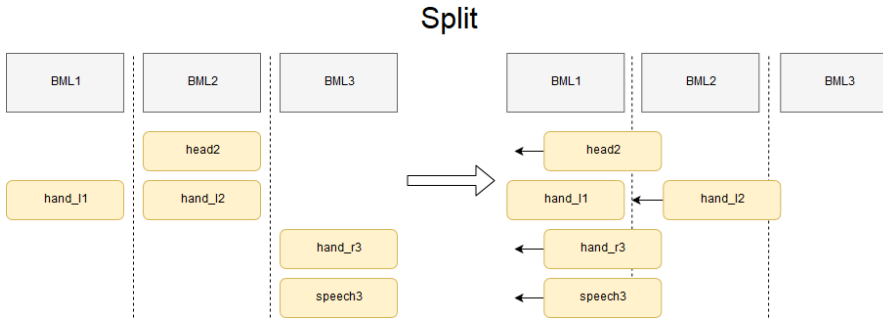


Рис. 9. Режим split: теги head2, hand_r3, speech3 могут выполниться немедленно

Режим join объединяет теги BML-пакета и они вместе конкурируют в очереди за возможность занять необходимые исполнительные органы робота. Обработка тегов из пакета начинается, как только освободятся все из указанных в пакете исполнительных органов. На **Рисунке 10** изображена последовательность выполнения тегов BML в очереди по порядку, а справа — очередь робота, в которой каждому из трех пакетов установлен режим join.

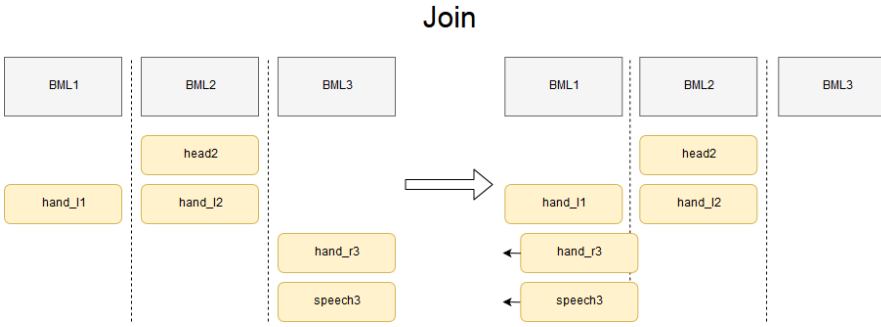


Рис. 10. Режим join: теги BML3 могут выполниться немедленно, теги BML2 ожидают завершения BML1

3.2. Синхронизация тегов BML

Архитектура робота позволяет синхронизировать теги внутри одного пакета BML. При конвертации жеста в базу в обязательном порядке, наряду с метками начала start и конца end, пиковой точке жеста приписывается временная метка stroke, которая необходима для синхронизации между высказыванием и жестом или различными элементами жеста (**Рис. 11**). Е. А. Гришина называет такую точку основной семантической частью жеста, отмечает, что именно в эту точку «все параметры, характерные для данного жеста, достигают максимального уровня напряженности» [Grishina, 2017, с. 23]. Автор также подчеркивает, что эта фаза жеста обязательна и может быть осуществлена самостоятельно, без предшествующих и последующих стадий. Такая точка важна для маховых жестов и кивков, обслуживающих такие функции как *апелляция, эмфаза, пренебрежение* и др.

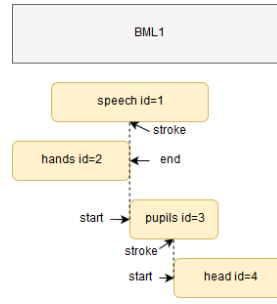


Рис. 11. Синхронизация тэгов внутри пакета BML

Синхронизация внутри пакета обеспечивается за счет уникального идентификатора тега и временных меток жеста (start, stroke, end). Уникальный идентификатор помогает точно определить тег, с которым необходимо синхронизироваться, а временная метка — указывает, какой именно момент выполнения жеста должен совпасть у тегов.

Перед выполнением теги, которые необходимо синхронизировать, выравниваются по времени, так чтобы временная метка одного тега выполнялась одновременно с другой. Для этого система откладывает начало выполнения тега, у которого метка наступает раньше, на необходимое для синхронизации время. Вне зависимости от режима, установленного в пакете BML, группа синхронизированных тегов рассматривается в режиме join, для того чтобы гарантировать исполнение описанных жестовых движений вместе — по указанным точкам синхронизации.

4. Заключение

Описанная архитектура позволяет моделировать на роботе ключевые особенности поведения информантов в корпусе. В частности, данная архитектура позволяет генерировать разнообразное коммуникативное поведение: (а) за счет использования различных способов выражения какой-либо коммуникативной функции; (б) за счет выражения определенной коммуникативной функции с помощью одного или нескольких исполнительных органов; (в) за счет гибкого комбинирования жестов с разными коммуникативными функциями.

Литература

1. *Beuter N., Spexard T., Lutkebohle I., Peltason J., Kummert F.* (2008), Where is this? — gesture based multimodal interaction with an anthropomorphic robot, 8th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2008), pp. 585–591. DOI: 10.1109/ICHR.2008.4756009
2. *Breazeal C., Scassellati B.* (2002), Robots that imitate humans, Trends in Cognitive Sciences. V. 6(11), pp. 481–487.
3. *Brugman H., Russel A.* (2004), Annotating Multimedia Multi-modal resources with ELAN, Proceedings of the 4th International Conference on Language Resources and Language Evaluation (LREC 2004), pp. 2065–2068.
4. *ELAN* (Version 5.0.0) [Computer software]. Nijmegen: Max Planck Institute for Psycholinguistics. URL: https://tla.mpi.nl/tools/tla-tools/elan/ (дата обращения 01.11.2017).
5. *Grishina E. A.* (2017) Russian gestures from a linguistic perspective (A collection of corpus studies) [Russkaya zhestikulyaciya s lingvisticheskoj tochki zreniya (Korpusnye issledovaniya)], LRC Publishing House, Languages of Slavic Culture, M.
6. *Kirbya R., Forlizzib J.* (2010), Simmons R. Affective social robots, Robotics and Autonomous Systems. V. 58(3), pp. 322–332. DOI: 10.1016/j.robot.2009.09.015

7. *Kotov A. A.* (2012), "Orwell machine":approaches to the automatic generation of influential texts ["Mashina Orujella": podhody k avtomaticheskomu sozdaniju vozdejstvujushhih tekstov] // Understandig in a Communication [Ponimanie v kommunikacii], Vol. 1., Yaroslavl': Univ. of Yaroslavl'.
8. *Kotov A., Arinkin N., Zaidelman L., Zinina A.* (in press), Linguistic approaches to robotics: from text analysis to the synthesis of behavior.
9. *Kotov A. A., Zinina A. A.* (2015a), Functional annotation of communicative actions in REC corpus [Funkcional'naja razmetka kommunikativnyh dejstvij v korpuse "REC"] // Corpora Linguistics — 2015 [Trudy mezhdunarodnoj konferencii "Korpusnaja lingvistika — 2015"], SPb: Univ. of SPb, pp. 287–295.
10. *Kotov A. A., Zinina A. A.* (2015b), Functional analysis of nonverbal communicative behavior [Funkcional'nyj analiz neverbal'nogogo kommunikativnogo povedeniya], Computational Linguistics and Intellectual Technologies [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii], Issue. 14., vol. 1, pp. 299–310. RSUH, M.
11. *Kotov A., Zinina A., Filatov A.* (2015), Semantic Parser for Sentiment Analysis and the Emotional Computer Agents. Proceedings of the AINL-ISMW FRUCT 2015, pp. 167–170.
12. *Kotov A., Budyanskaya E.* (2012), The Russian Emotional Corpus: Communication in Natural Emotional Situations, Computational Linguistics and Intellectual Technologies [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii], Issue 11 (18). vol. 1, pp. 296–306. RSUH, M.
13. *Kotov A. A.* (2003), Mechanisms of speech influence in mass media texts [Mehanizmy rechevogo vozdejstvija v publicisticheskih tekstah SMI], PhD in Philology, M.
14. *Lee D., Ahn H. S., Choi J. Y.* (2009), A general behavior generation module for emotional robots using unit behavior combination method, The 18th IEEE International Symposium on Robot and Human Interactive Communication, pp. 375–380. DOI: 10.1109 / ICSMC.1995.538385
15. *Leite I., Pereira A., Martinho C., Paiva A.* (2008), Are Emotional Robots More Fun to Play With? Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, Technische Universität München, Munich, Germany, August 1–3, pp. 77–82.
16. *Velásquez J. D.* (1998), When Robots Weep: Emotional Memories and Decision-Making, AAAI-98 Proceedings. Copyright, pp. 70–75.
17. *Vilhjálmsson H., Cantelmo N., Cassell J., Chafai N., Kipp M., Kopp S., Mancini M., Marsella S., Marshall A., Pelachaud C., Ruttkay Z., Thórisson K., van Welbergen H., van der Werf R.* (2007), The Behavior Markup Language, Recent Developments and Challenges in Intelligent Virtual Agents, pp. 99–111.

Abstracts

INTRA-TEXT COHERENCE AS A MEASURE OF TOPIC MODELS' INTERPRETABILITY

Alekseev V. A. (wasya.alekseev@gmail.com), **Bulatov V. G.** (bt.uytya@gmail.com), **Vorontsov K. V.** (vokov@forecsys.ru), Moscow Institute of Physics and Technology (State University)

The article is devoted to the problem of how to automatically measure the interpretability of topic models. Some new, intra-text, approaches to estimate the interpretability of the topics are proposed. Computational experiments are conducted with the use of text files from "Post-Nauka", which is a collection of popular science content.

IMPROVING PART-OF-SPEECH TAGGING VIA MULTI-TASK LEARNING AND CHARACTER-LEVEL WORD REPRESENTATIONS

Anastasyev D. G. (daniil_an@abbyy.com), **Gusev I. O.** (ilya.gusev@phystech.edu), **Indenbom E. M.** (eugene_i@abbyy.com), ABBYY, Moscow Institute of Physics and Technology, Moscow, Russia

In this paper, we explore the ways to improve POS-tagging using various types of auxiliary losses and different word representations. As a baseline, we utilized a BiLSTM tagger, which is able to achieve state-of-the-art results on the sequence labelling tasks. We developed a new method for character-level word representation using feedforward neural network. Such representation gave us better results in terms of speed and performance of the model. We also applied a novel technique of pretraining such word representations with existing word vectors. Finally, we designed a new variant of auxiliary loss for sequence labelling tasks: an additional prediction of the neighbour labels. Such loss forces a model to learn the dependencies inside a sequence of labels and accelerates the process of training. We test these methods on English and Russian languages.

DISCOVERING DIALECTAL DIFFERENCES BASED ON ORAL CORPORA

Andriyanets V. (blindedbysunshine@gmail.com), **Daniel M.** (misha.daniel@gmail.com), International Linguistic Convergence Laboratory, NRU HSE, Moscow, Russia; **Pakendor B.** (brigitte.pakendorf@cnrs.fr), Laboratoire "Dynamique du Langage", UMR5596, CNRS & Université de Lyon, Lyon, France

This paper discusses a method to detect statistically significant linguistic differences between corpora while factoring in possible variability within the very corpora to be compared. Specifically, we compare two small corpora of dialects of Even, Bystraja and Lamunkhin Even, in an attempt to identify morphemes that are more frequent in either of the corpora. To investigate whether this difference might be due to an over-representation of a speaker who happens to be an outlier in terms of using a particular morpheme, we use DP, a measurement of evenness of the distribution of a specific linguistic feature across subcorpora of the same corpus.

RUSSIAN CONSTRUCTIONS CHAINIK DOLGO (NE) ZAKIPAET, KOMP'IUTER DOLGO (NE) ZAGRUZHAETSIA…

Apresjan V. Ju. (valentina.apresjan@gmail.com)[1,2], **Shmelev A. D.** (shmelev.alexei@gmail.com)[2–4], [1]National Research University Higher School of Economics; [2]Vinogradov Institute of Russian Language, Russian Academy of Sciences; [3]Moscow Pedagogical State University; [4]St Tikhon's Orthodox University

The paper deals with a curious phenomenon of quasi-synonymy that occurs in Russian between sentences with non-negated and negated predicates in the construction with the adverb dolgo 'for a long time'. Consider sentences like Chainik dolgo zakipal 'It took the kettle a long time to boil, lit. Kettle for a long time boiled' vs. Chainik dolgo ne zakipal 'It took the kettle a long time

to boil, lit. Kettle for a long time not boiled'. The paper is an attempt to define the semantic and pragmatic mechanisms of such quasi-synonymy, as well as semantic and aspectual classes of predicates where it occurs. It also considers subtle semantic, pragmatic and communicative differences associated with non-negated and negated construction, respectively.

Such quasi-synonymy occurs primarily in cases when the predicate belongs to the aspectual class of accomplishments and denotes a telic process or action with a desired result ('to boil', 'to cool down', 'to warm up', 'to grow up', 'to finish', etc.). Those predicates include two major semantic components, that is, a lasting process or action and an instant result. In the imperfective aspect they allow at least two possible interpretations, namely, of a process and that of a result. Similar interpretations of sentences with such predicates occur due to different scope assignments of negation and dolgo. In sentences with non-negated predicate dolgo has scope over the 'process' component in the verb; in sentences with negated predicate negation has scope over the 'result' component of the verb while at the same time falling into the scope of dolgo. The former type of sentences describes long-lasting processes, whereas the latter type describes long-awaited results, which pragmatically amount to the same thing.

DISAMBIGUATION OF SCOPE IN WRITTEN ENGLISH TEXTS

Apresyan V. Ju. (valentina.apresjan@gmail.com, vapresyan@hse.ru), National Research University "Higher School of Economics", Moscow, Russia

The paper is a corpus study of the factors involved in disambiguating potential scope ambiguity in written sentences with negation and universal quantifier all, such as I cannot visit all these universities, which, depending on topic-focus assignment, can alternatively mean 'I cannot visit any of these universities' (cannot is focus) and 'I cannot visit some of these universities' (all is focus). The factors at play in scope disambiguation are the syntactic function of the constituent containing all (subject, direct complement, adjunct); the status of the main predicate and all with respect to the information structure of the utterance (topic vs. focus); veridical vs. non-veridical context; sentence type (unreal conditional, rhetorical question); and pragmatic implicatures pertaining to the situations described in the utterances. The paper also demonstrates differences in the frequency distribution of various scope readings and their underlying causes, as well as formulating typical contexts for each scope interpretation.

HOW MUCH DOES A WORD WEIGH? WEIGHTING WORD EMBEDDINGS FOR WORD SENSE INDUCTION

Arefyev N. (narefyev@cs.msu.ru), Lomonosov Moscow State University & Samsung Moscow Research Center, Moscow, Russia;
Ermolaev P. (permolaev@cs.msu.ru), Lomonosov Moscow State University, Moscow, Russia;
Panchenko A. (panchenko@informatik.uni-hamburg.de), University of Hamburg, Hamburg, Germany

The paper describes our participation in the first shared task on word sense induction and disambiguation for the Russian language RUSSE'2018 (Panchenko et al., 2018). For each of several dozens of ambiguous words, the participants were asked to group text fragments containing it according to the senses of this word, which were not provided beforehand, therefore the „induction" part of the task. For instance, a word "bank" and a set of text fragments (also known as "contexts") in which this word occurs, e.g. "bank is a financial institution that accepts deposits" and "river bank is a slope beside a body of water" were given. A participant was asked to cluster such contexts in the unknown in advance number of clusters corresponding to, in this case, the "company" and the "area" senses of the word "bank". The organizers proposed three evaluation datasets of varying complexity and text genres based respectively on texts of Wikipedia, Web pages, and a dictionary of the Russian language.

We present two experiments: a positive and a negative one, based respectively on clustering of contexts represented as a weighted average of word embeddings and on machine translation using two state-of-the-art production neural machine translation systems. Our team showed the second best result on two datasets and the third best result on the remaining one dataset among 18 participating teams. We managed to substantially outperform competitive state-of-the-art baselines from the previous years based on sense embeddings.

MORPHOLOGICAL SEGMENTATION WITH SEQUENCE TO SEQUENCE NEURAL NETWORK

Arefyev N. V. (narefjev@cs.msu.su), Lomonosov Moscow State University, Moscow, Russia; Samsung Moscow Research Center, Moscow, Russia; **Gratsianova T. Y.** (tgratsianova@cs.msu.su), **Popov K. P.** (kpopov94@ya.ru), Lomonosov Moscow State University, Moscow, Russia

Morphological segmentation is an important task of natural language processing as it can significantly improve the processing of unfamiliar and rare words in different tasks that involve text data. In this paper we present datasets in English and Russian for learning and evaluating morphological segmentation algorithms, demonstrate the method based on the sequence to sequence neural model and show that the proposed approach shows better results in comparison with other existing methods of morpheme segmentation. We start from an English dataset, which is already available and only minor preprocessing has been made, and then we experiment with the Russian language, where we could not obtain prepared data. So, some more serious preprocessing issues are included. Moreover, we demonstrate how morphological segmentation can improve another natural language processing task—evaluation of words semantic similarity. To achieve this goal, first we try to reproduce the best results of the participants of Russian words semantic similarity competition (RUSSE), which was conducted in Dialogue 2015 conference. Then we show how with the help of smart morpheme segmentation these results can be advanced.

FRAMEWORK FOR RUSSIAN PLAGIARISM DETECTION USING SENTENCE EMBEDDING SIMILARITY AND NEGATIVE SAMPLING

Belyy A. V. (anton.belyy@gmail.com)[1,2], **Dubova M. A.** (marina.dubova.97@gmail.com)[3], [1]ITMO University, Saint Petersburg, Russia; [2]B Tochka Bank QIWI Bank (JSC), Yekaterinburg, Russia; [3]Saint Petersburg State University, Saint Petersburg, Russia

In this paper, we propose a new approach for advanced plagiarism detection in Russian language. It is based on a classifier, dealing with two different types of sentence similarity measures: token set similarity and cosine similarity between sentence embeddings (based on pretrained RusVectōrēs, unsupervised fastText, and supervised StarSpace models). The diversity of feature space makes it possible to detect different types of plagiarism, starting from simple copy&paste cases and ending with complex manual paraphrases. The proposed approach implies an ability to focus on the particular plagiarism type identification, allowing to train a universal model at the same time. The method shows great results on detection of different types of plagiarism and outperforms the previous approach.

QUALITY EVALUATION AND IMPROVEMENT FOR HIERARCHICAL TOPIC MODELING

Belyy A. V. (anton.belyy@gmail.com), ITMO University, Saint Petersburg, Russia; B Tochka Bank QIWI Bank (JSC), Yekaterinburg, Russia;
Seleznova M. S. (maria.selezniova@phystech.edu), **Sholokhov A. K.** (ak.sholokhov@gmail.com), **Vorontsov K. V.** (vokov@forecsys.ru), Moscow Institute of Physics and Technology (State University),Moscow, Russia

Generic topics of large-scale document collections can often be divided into more specific subtopics. Topic hierarchies provide a model for such topic relation structure. These models can be especially useful for exploratory search systems. Various approaches to building hierarchical topic models have been proposed so far. However, there is no agreement on a standard approach, largely due to the lack of quality metrics to compare existing models. To bridge this gap we propose automated evaluation metrics which measure the quality of topic-subtopic relations (edges) of a topic hierarchy. We compare automated evaluations with human assessment to validate the proposed metrics. Finally, we show how the proposed metrics can be used to control and to improve the quality of existing hierarchical models.

SEMANTIC ANALYSIS WITH INFERENCE: HIGH SPOTS OF THE FOOTBALL MATCH

Boguslavsky I. M. (bogus@iitp.ru)[1,2], **Frolova T. I.** (tfrolova@gmail.com)[1], **Iomdin L. L.** (iomdin@gmail.com)[1], **Lazursky A. V.** (lazursky@mail.ru)[1], **Rygaev I. P.** (irygaev@gmail.com)[1], **Timoshenko S. P.** (nyrestein@gmail.com)[1], [1]A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia, [2]Universidad Politécnica de Madrid, Madrid, Spain

The paper describes a new version of the semantic analyzer SemETAP. Our approach is based on the assumption that the depth of understanding is growing with the number of inferences we can draw from the text. The salient features of SemETAP include: 1) intensive use of both linguistic and background knowledge. The former is incorporated in the Combinatorial Dictionary and the Grammar, and the latter is stored in the Ontology and Repository of Individuals. 2) Words and concepts of the ontology may be supplied with explicit decompositions for inference purposes. 3) Two levels of semantic structure are distinguished. Basic semantic structure (BSemS) interprets the text in terms of ontological elements. Enhanced semantic structure (EnSemS) extends BSemS by means of a series of inferences. 4) A new logical formalism Etalog is developed in which all inference rules are written. Semantic analysis with inference allows us to extract implicit information. The analyzer is tested on the task of interpreting high spots of the football match.

TERM EXTRACTION FOR CONSTRUCTING SUBJECT INDEX OF EDUCATIONAL SCIENTIFIC TEXT

Bolshakova E. I. (eibolshakova@gmail.com), Moscow State Lomonosov University, National Research University Higher School of Economics, Moscow, Russia; **Ivanov K. M.** (ivanov.kir.m@yandex.ru), Moscow State Lomonosov University

Subject index, or back-of-the-book index, is a device intended to provide an easy access to relevant fragments of a text document. Subject indexes usually contain particular single-word and multi-word terms from the corresponding documents. Such indexes are especially useful for reading large documents with specialized terminology, as well as educational texts in difficult scientific and technical areas. The central problem of back-of-the-book indexing is recognition of terms to be included into the index. The paper describes a method developed for extracting and filtering terms from a given educational scientific text, with the purpose of reliable term selection in computer indexing systems. The method is primarily based on rules with lexico-syntactic patterns representing linguistic information about terms and typical contexts of their usage in Russian scientific and educational texts; simple occurrences statistics of terms is used as well. Experimental evaluation of the method has shown a considerable increase of precision and recall of term extraction compared with the widely-used standard techniques.

USING MACHINE TRANSLATION FOR AUTOMATIC GENRE CLASSIFICATION IN ARABIC

Bulygin M. V. (bulyginmv1996@gmail.com)[1], **Sharoff S. A.** (s.sharoff@leeds.ac.uk)[1,2]; [1]Russian State University of Humanities, Moscow, Russia, [2]Leeds University, Leeds, UK

This paper addresses the task of automatic genre classification for Arabic within the Functional Text Dimensions framework, which allows texts to get a reliable genre description, while maintaining an adequate amount of genre labels. Our aim in this study is to build an automatic classification model that can annotate any Web text in Standard Arabic in terms of genres. To build the training corpus we translated English and Russian annotated texts into Arabic using Google MT. For building the model experimented with various machine learning approaches, such as Logistic Regression, SVM, LSTM, and different features, such as words, character n-grams and embedding vectors. For testing the classification models, we collected and annotated in terms of FTDs our own corpus of Arabic Web texts. The best performing model offers reasonable classification accuracy in spite of being based on a training corpus produced by MT.

BOUNDARY EXPRESSION IN VERBS AND GESTURE: DIFFERENCES BETWEEN L1 AND L2 SPEAKERS

Denisova V. A. (valeriia.deni@gmail.com)[1,2], **Cienki A.** (a.cienki@vu.nl)[1,2], **Iriskhanova O. K.** (iriskhanova@me.com)[1]; [1]Moscow State Linguistic University, Moscow, Russia; [2]Vrij Universiteit Amsterdam, Amsterdam, the Netherlands

The notion of event boundaries is closely connected with the category of aspect. Aspectual forms show different views of "internal temporal consistuency of a situation" (Comrie 1976:3) and, consequently, construals of events in different ways. Recently scholars have started looking into the core of the aspectual distinction through multimodality, considering hand gestures. On the basis of Russian and French oral narratives produced by native speakers, we conducted a study, testing our hypothesis about the existence of direct correlation between the expression of boundaries in verbs and in gestures. Means of boundary expression regarded for Russian on the verbal level were perfective (soveršennyj vid) and imperfective (nesoveršennyj vid) verbs, and for French—passé composé and imparfait. On the kinesthetic level we distinguished between bounded gestures (i.e., involving a pulse of movement) and unbounded gestures (i.e., smooth by nature). While for French L1 we found a direct correlation between gesture boundary schemas and aspectual forms, the results for Russian L1 did not support our hypothesis. With a view to these differences between the two languages, we studied the boundedness correlation in oral narratives produced by Russians speaking French as L2 (CEFR levels B2-C1). The comparison between L1 and L2 narratives revealed a certain change of gestural patterns: the Russian speakers of French L2 used almost the same number of unbounded and bounded gestures with the perfective verb forms and more unbounded gestures with the imperfective forms, thus moving closer towards French L1 speakers' verb-gesture patterns. The use of gestures can be accounted for by a series of noise factors related to language peculiarities, the cognitive mechanism of profiling and challenges of speaking in L2.

GERMAN CONSTRUCTIONS WITH MODAL VERBS AND THEIR RUSSIAN CORRELATES: A SUPRACORPORA DATABASE PROJECT

Dobrovol'skij D. O. (dobrovolskij@gmail.com), Russian Language Institute of the RAS, Moscow, Russia; Zalizniak **Anna A.** (anna.zalizniak@gmail.com), Institute of Linguistics of the RAS; Institute of Informatics Problems of the FRC CSC RAS, Moscow, Russia

The paper outlines the principles of analyzing German and Russian modal constructions. Our first task is to clarify the set of meanings of German modal verbs and the conditions for their implementation. The second task is to describe the means of expressing modal values in Russian that are encountered in parallel corpora as functional equivalents of constructions with German modal verbs. As empirical data we use a representative array of parallel German-Russian texts from the Russian National Corpus (RNC). A supracorpora database of translation correspondences is constructed, in which both the German constructions with modal verbs and their Russian translation equivalents are attributed an annotation of their relevant characteristics. This database, on the one hand, is a valuable linguistic resource that can be used, among other things, to create a new generation of electronic interactive German-Russian and Russian-German dictionaries. On the other hand, the inventory of Russian construction types with (implicit) modal meanings constructed on this database will contribute to the Construction Grammar and confirm the continuity between grammar and lexicon.

DISCOURSE MARKER 'TIPA' ACCORDING TO THE DATA OF RUSSIAN NATIONAL CORPUS: ITS ORIGIN, SEMANTICS AND PRAGMATICS

Egorova M. A. (drajenka@gmail.com), RSUH, Moscow, Russia

Discourse marker tipa became widespread in colloquial Russian in the decade 1990s–2000s. However, until recently, it has gained little attention. In this paper we use the data from the Russian National Corpus and we aim to accomplish the following goals: 1) to highlight the origin of the discourse marker tipa from the noun tip 'type', 2) to describe the semantics of the discourse marker tipa as well as that of the partly grammaticalized element tipa as part of parametric constructions. We base our approach mainly on the results achieved by Susanne Fleischman and Marina Yaguello.

A STUDY OF MACHINE LEARNING ALGORITHMS APPLIED TO GIS QUERIES SPELLING CORRECTION

Fomin V. V. (wadimiusz@gmail.com), Novosibirsk State University, Novosibirsk, Russia;
Bondarenko I. Yu. (i.yu.bondarenko@gmail.com), [2]GIS, Novosibirsk, Russia

The problem of spelling correction is crucial for search engines as misspellings have a negative effect on their performance. It gets even harder when search queries are related to a specific area not quite covered by standard spell checkers, such as geographic information systems (GIS). Moreover, standard spell-checkers are interactive, i.e. they can notice a misspelled word and suggest candidate corrections, but picking one of them is up to the user. This is why we decided to develop a spelling correction unit for 2GIS, a cartographic search company. To do this, we have extracted and manually annotated a corpus of GIS lookup queries, trained a language model, performed various experiments to find the best feature extractor, then fitted a logistic regression using an approach suggested in SpellRuEval, and then used it iteratively to get a better result. We have then measured the resulting performance by means of cross-validation, compared at against a baseline and observed a substantial increase. We also present an interpretation of the result achieved by calculating and discussing the importance of specific features and analyzing the output of the model.

DISCOVERING AND ASSESSING HEATED ARGUMENTS AT THE DISCOURSE LEVEL

Galitsky B. (boris.galitsky@oracle.com)[1,2], **Taylor R.** (ray.taylor@oracle.com)[1];
1Oracle Inc., Redwood Shores CA, USA, [2]Higher School of Economics, Moscow, Russia

The problem of detecting heated arguments in text such as political debates and customer complaints is formulated as tree kernel learning of discourse structures. Affective argumentation structure is discovered in the form of discourse trees extended with edge labels for communicative actions. Extracted argumentation structures are then encoded as defeasible logic programs and are subject to dialectical analysis, to establish the validity of the main claim being communicated. We evaluate the accuracy of each step of this affect processing pipeline as well as overall performance.

THE INFLUENCE OF SYNTAX ON PROSODY: THE EXPERIMENTAL DATA FROM A STUDY OF ONE RUSSIAN TEXT

Grashchenkov P. V. (pavel.gra@gmail.com), Lomonosov Moscow State University; IOS RAS, MSUE; Moscow, Russia; **Kirillova A. A.** (anastasya_kirillova@hotmail.com), **Smirnova O. S.** (kisaolga@mail.ru), Lomonosov Moscow State University; Moscow, Russia

The paper examines dependencies between the syntactic and prosodic structure with particular attention to the pausation and different levels of prosodic boundary strength. The research is based on the prosodic data markup for a spoken Russian text and the manual tagging of this text with the relevant syntactic constituent boundaries. Two types of structures, the finite clause and the asyndetic coordination, exhibit a strong positive correlation with the appearance of a pause and the perceptual prosodic boundary. We also demonstrate the presence of a substantial correlation between the syntactic embedding depth and prosodic boundaries. The results of our research show a significant connection between some of the initially proposed syntactic factors and prosodic structure. We thus anticipate that prosodic modules of TTS systems can benefit from taking certain syntactic information into consideration.

SUPRACORPORA DATABASE AS AN INSTRUMENT OF THE STUDY OF THE FORMAL VARIABILITY OF CONNECTIVES

Inkova O. Yu. (Olga.Inkova@unige.ch), Institute of Informatics Problems, FRC CSC RAS, Moscow, Russia; University of Geneva, Geneva, Switzerland

The article intends to describe the formal variation of the connectors of the Russian language on the basis of a cognitive-semantic approach. Every discourse variant DV of a connector K, i.e. the specific form assumed by K in a discourse section, is singled out, and registered in the

supracorpora database of connectors (SCDB), in which a system of intersecting clusters has been developed, allowing to assign in the course of the annotation the same DV to different structural clusters. In the next phase, on the base of further semantic analysis, the DVs with a common element are combined into a structural-semantic complex around a basic form: the minimal linguistic unit that enables the speaker to express a certain logical-semantic relation, and the listener to identify it. In conclusion, criteria for describing the formal variation of the connectors are proposed, as well as examples of the "profiles" of the basic forms. They reflect the potential of linguistic means that the speaker has at his disposal to express one or another logical-semantic relations or one of their combinations.

TO WHAT EXTENT IS THE CONJUNCTION 'KHOTYA' LANGUAGE-SPECIFIC?

Inkova O. Yu. (Olga.Inkova@unige.ch), Institute of Informatics Problems, FRC CSC RAS, Moscow, Russia; University of Geneva, Geneva, Switzerland; **Nuriev V. A.** (nurieff.v@gmail.com), Institute of Linguistics, RAS, Moscow, Russia; Institute of Informatics Problems, FRC CSC RAS, Moscow, Russia

The paper describes the Russian connective khotya ('although') from a contrastive perspective. First, it focuses on the semantic description of the connective and proposes to differentiate its four meanings, namely, concessive propositional, concessive illocutionary, adversative propositional and adversative illocutionary. The paper analyzes the functioning of the connective khotya (prototypical marker of concessive relations) and that of the connective no ('but', prototypical marker of adversative relations). In so doing, it comes to the following conclusion: the adversative meaning of khotya develops on the basis of its concessive meaning as the connection between the situations presented in the textual fragments that are linked by the connective becomes less logical. Similarly, i. e. vice-versa, as the logical connection between situations becomes stronger, this gives rise to a concessive interpretation in utterances with no. Further, the paper takes a closer look at French equivalents khotya gets, when occurring in each of its four meanings. The concluding section attempts to define the degree of language-specificity of khotya. To this end, several parameters are considered: (1) cases where the connective has a zero equivalent, (2) cases of divergent translation (the connective is translated by a non-connective), (3) number of translation patterns. To perform a contrastive analysis and to collect statistical data, the supracorpora database of connectives is used. The database is built upon the parallel Russian-French and French-Russian subcorpora of the RNC.

ONCE AGAIN ON MICROSYNTACTIC CONSTRUCTIONS FORMED WITH FUNCTIONAL WORDS: 'TO I DELO' 'EVERY NOW AND THEN'

Iomdin L. L. (iomdin@iitp.ru), Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Moscow, Russia

The paper continues a series of research studies into the microsyntax of Russian, conducted by the author for a considerable period of time. Specifically, the focus is on the adverbial syntactic idiom to i delo '≈ every now and then', which seems very interesting and instructive as it combines implicit semantic features and a unique set of syntactic facets that could be revealed by both present-day and diachronic linguistic data. This syntactic idiom is considered against the background of other microsyntactic elements that happen to be its neighbors in the dictionary but feature a substantially different set of linguistically relevant properties. It is shown how phraseological units of such kind can be presented in the Microsyntactic dictionary of Russian, under development by the author and his colleagues, and in the corpus of texts annotated with microsyntactic phenomena.

EFFICIENCY OF TEXT READABILITY FEATURES IN RUSSIAN ACADEMIC TEXTS

Ivanov V. V. (nomemm@gmail.com), Innopolis University, Innopolis, Russia;
Solnyshkina M. I. (mesoln@yandex.ru), **Solovyev V. D.** (maki.solovyev@mail.ru), Kazan Federal University, Kazan, Russia

This paper addresses the problem of readability assessment for Russian texts and investigates the impact of 24 lexical, syntactic and frequency features. The research was conducted on Russian Readability Corpus containing two sub-corpora, two sets of 5–11 grade level textbooks on Social studies for native speakers of Russian. The sub-corpora were collected for research purposes, annotated and marked as BOG and NIK. The application of the Ridge regression has demonstrated the connection between readability and average sentence length, average number of coordinating chains, average number of sub-trees, frequency and lexical features. The results of the study have the potential to be applied in a wide variety of areas including primarily education, as well as webpage design, document management.

CORPUS-BASED INVESTIGATION OF QUOTATION IN RUSSIAN SIGN LANGUAGE

Khristoforova E. A. (evkhristoforova@gmail.com), **Kimmelman V. I.** (vadim.kimmelman@gmail.com), Russian State University for the Humanities, Moscow, Russia

This paper presents corpus-based research of quotation constructions in Russian Sign Language (RSL). Quotation constructions have been observed from different perspective in different signed and spoken languages [Brendel, Meibauer, Steinbach 2011]; [Litvinenko et al. 2009]. Based on the corpus of spontaneous narratives recorded from RSL signers [Burkova 2015], we conducted a quantitative analysis of these constructions. We analyzed constituents of quotation construction, such as the source (author of utterance) indication, the introducing matrix predicate, and the quote. Our investigation of non-manual markers in the corpus revealed that non-manual marking of quotation is optional for RSL quotations. We distinguished direct and indirect quotations in our data based on the reference of indexical elements, the use of subordinating conjunction, and the imperative mood. We found that in RSL non-manuals do not mark the direct/indirect type of quotation. Our data show that RSL signers tend to use direct quotation much more frequently than indirect quotation. In addition, we compared our findings with the data on quotation constructions in some other sign languages and with the studies of quotation in natural discourse of spoken languages. This comparison showed that RSL quotations share core properties with quotations in spoken and signed languages [Litvinenko et al. 2009].

LANGUAGE PRODUCTION AND COMPREHENSION IN FACE-TO-FACE MULTICHANNEL COMMUNICATION

Kibrik A. A. (aakibrik@gmail.com), **Fedorova O. V.** (olga.fedorova@msu.ru), Institute of Linguistics RAS and Lomonosov Moscow State University, Moscow, Russia

Although language production and comprehension are parts of one and the same linguistic capacity, they have been studied separately for a long time. A key issue in the present day research is how the two processes are related, and whether transitions from thought to language and vice versa are accomplished by a single or two separate systems. Important progress in this area has been achieved in the field of psycho- and neurolinguistics; a brief review is provided in Section 1. In this paper we explore the production—comprehension relationship on the basis of our multichannel resource "Russian Pear Chats and Stories". In Section 2 we describe this resource, including the stimulus material, data collection setup, participants and corpus size, and technical aspects. Section 3 lays out two main theoretical notions: a model of face-to-face multichannel communication and a scheme of the production-comprehension interweaving in each interlocutor. In subsequent sections we discuss three case studies of production—comprehension relationships: relative contributions of kinetic channels to discourse understanding (Section 4), turn-taking and eye gaze (Section 5), and multichannel continuity (Section 6). The evidence of the multichannel corpus suggests a cognitive architecture that integrates language production and comprehension.

CREATING A CORPUS OF SYNTACTIC CO-OCCURRENCES FOR RUSSIAN

Klyshinsky E. S. (klyshinsky@mail.ru), Keldysh IAM RAS, Moscow, Russia;
Lukashevich N. Y. (natalukashevich@mail.ru), **Kobozeva I. M.** (kobozeva@list.ru),
Moscow State University, Moscow, Russia

In the paper we discuss methods used to create CoSyCo, a corpus of syntactic co-occurrences, which provides information on syntactically related words in Russian. We describe a list of shallow parsing templates, which were used to collect data for CoSyCo. The paper includes an overview of the corpora collected for CoSyCo creation and an outline of how the noun 'virus' is used in its subcorpora as an example of the information which can be obtained from this online resource.

LEARNING WORD EMBEDDINGS FOR LOW RESOURCE LANGUAGES: THE CASE OF BURYAT

Konovalov V. P. (vaskonov@yahoo.ru), **Tumunbayarova Z. B.** (zhargal@zabgu.ru),
Transbaikal State University, Chita, Russia

Word-vector representations have been extensively studied for rich resource languages with large text datasets. However, only a few studies analyze semantic representations of low resource languages, when only small corpus is available. In this study we introduce a methodology and compare techniques to learn semantic representations of low resource languages. The proposed methodology consists of defining accurate preprocessing steps, applying language-independent stemmer and learning word-vector representations. In addition, we propose a simple word embeddings evaluation scheme that can be easily adapted to any language. By using this methodology we learn word-vector representations for Buryat language. In order to promote further research we make the source code and the resulting word embeddings corpus publicly available.

HOW INTONATION STRUCTURES SPOKEN NARRATIVES: NON-FINAL PHASE CONTEXTS

Korotaev N. A. (n_korotaev@hotmail.com), RSUH, RANEPA, Moscow, Russia

Topic—focus articulation in Russian has been mainly studied against isolated utterances. In a categorical sentence, this communicative opposition is reflected in the linear-accentual structure [Paducheva 2015]. For a simple declarative sentence, that would normally mean that the topic (theme) comes first and has a rising phrasal accent, while the focus (rheme) completes the utterance and is pronounced with a falling accent. At the same time, these formal features do more than just differentiate between topics and foci; they also mark the discourse-semantic category of phase [Kodzasov 2009]. In syntactically simple utterances, topics tend to correlate with anticipated continuation, hence non-final phase; foci are usually phase-final. As I intend to show in this paper, the non-final phase provides a variety of contexts that challenge the topic—focus distinction. The study is based on the "Stories about presents and skiing"—a collection of prosodically annotated spoken narratives.

In Section 1, I concentrate on issues within a simple clause, where non-final verbal elements often have a fuzzy communicative interpretation. In Section 2, I analyze complex syntactic structures. The data show that non-final clauses may demonstrate both thematic and rhematic properties with regard to their intonation patterns, internal structure and discourse function. Hence, one can claim that some non-final clauses are topics, while others are foci. However, a majority of non-final clauses in the analyzed corpus may not be unambiguously attributed to either of these categories. Section 3 provides a pilot study of complex intonation patterns. Only phase distinction being considered, utterances with more than one accentual phrase may follow either (i) the basic adaptation strategy (comprising a non-final rising accent and a final falling accent), or, more often, (ii) a complicated strategy: (a) multiple parallel adaption, (b) consecutive adaptation, or (c) parenthetical strategy.

FRAMES REVISITED: AUTOMATIC EXTRACTION OF SEMANTIC PATTERNS FROM A NATURAL TEXT

Kotov A. A. (kotov_aa@nrcki.ru), **Zaidelman L. Y.** (zaydelman_ly@nrcki.ru), **Arinkin N. A.** (arinkin_na@nrcki.ru), Kurchatov Institute; Russian State University for the Humanities; Moscow, Russia; **Zinina A. A.** (zinina_aa@nrcki.ru), Kurchatov Institute, Moscow, Russia; **Filatov A. A.** (alexander.filatov@hp.com), HP Inc., Moscow, Russia

Our project aims to design a syntactic parser, which constructs a semantic representation in a frame format: a clause is represented as a table of valencies, filled in with semantic markers. This representation is compared to a list of scripts—used to disambiguate and classify the semantic representation as well as to select an appropriate reaction for a companion robot F-2.

A DATABASE OF WORDBREAKS DISCURSIVE FEATURES IN RUSSIAN ORAL SPEECH: THE STRUCTURE, COMPOSITION AND APPLICATION

Krivnova O. F. (okrivnova@mail.ru), **Smirnova O. S.** (kisaolga@mail.ru), Moscow State University, Moscow, Russia

The paper discusses the most important results of the project “Hierarchy of prosodic phrasing in spoken language: controlling factors and means of realization”. The project was aimed at expanding the empirical base of phrasal prosody researches, which inadequacy is marked in many scientific areas: discourse theory, syntax, intonational phonology, general phonetics, speech synthesis and recognition etc. The introduction provides a brief description of the study background and formulates the tasks which were necessary to solve for the ultimate goal of the project planned for 3 years of implementation. The first section describes the characteristics of speech corpora created in the the project for construction of a complex, linguistic-prosodic database required for the study and modeling of prosodic phrasing in Russian speech, which takes into account, if possible, all controlling factors and means of realization. The second section is devoted to the description of the structure and composition of wordbreaks’ discursive features database (BDF), obtained on the basis of annotated, prosodically graduated and acoustically analyzed speech corpora. It should be noted the universality and flexibility of the format and structure of the database as a computer resource, freely admitting to extend its feature set and to detail their parametric characteristics. The third section illustrates as the BDF application for theoretical and statistical modelling of inter-level correlations “syntax—linguistic prosody” in both directions and “linguistic prosody and speech signal (acoustic speech)” in both directions. The conclusion summarizes the results of research and discusses some promising directions for further studies on relevant topics.

MENTAL PREDICATES IN METATEXT

Kustova G. I. (galinak03@gmail.com), Vinogradov Russian Language Institute of the Russian Academy of Sciences; Moscow State Pedagogical University; Moscow, Russia

The paper deals with metatext (parenthetical) constructions (MC) with mental verbs (znat’ ‘know’, ponimat’ ‘understand’, verit’ ‘believe’ and the like) in the 2nd person. The following problems are considered: is there a semantic correlation between the proposition and MC; what illocutionary function MC and proposition have. It was shown that some MCs are used only in interrogative sentences.

RUSSIAN WORD SENSE INDUCTION BY CLUSTERING AVERAGED WORD EMBEDDINGS

Kutuzov A. B. (andreku@ifi.uio.no), University of Oslo, Oslo, Norway

The paper reports our participation in the shared task on word sense induction and disambiguation for the Russian language (RUSSE’2018). Our team was ranked 2nd for the wiki-wiki dataset (containing mostly homonyms) and 3rd for the bts-rnc and active-dict datasets (containing mostly polysemous words) among all 19 participants.

The method we employed was extremely naive. It implied representing contexts of ambiguous words as averaged word embedding vectors, using off-the-shelf pre-trained distribu-

tional models. Then, these vector representations were clustered with mainstream clustering techniques, thus producing the groups corresponding to the ambiguous word' senses. As a side result, we show that word embedding models trained on small but balanced corpora can be superior to those trained on large but noisy data—not only in intrinsic evaluation, but also in downstream tasks like word sense induction.

AUTOMATED TEXT READABILITY ASSESSMENT FOR RUSSIAN SECOND LANGUAGE LEARNERS

Laposhina A. N. (antonina.laposhina@gmail.com),
Veselovskaya T. S. (tatianus2006@yahoo.com),
Lebedeva M. U. (m.u.lebedeva@gmail.com), **Kupreshchenko O. F.** (ofkupr@gmail.com),
Pushkin State Russian Language Institute (Moscow, Russia)

This paper presents an outline of the readability assessment system construction for the purposes of the Russian language learning. The system is designed to help educators easily obtain the information about the difficulty level of reading materials. The estimation task is posed here as a regression problem on data set of 600 texts and a range of lexico-semantic and morphological features. The scale choice and annotated text collection issues are also discussed. Finally, we present the results of the experiment with learners of Russian as a foreign language to evaluate the quality of a predictive model.

LEXICAL VARIATION: WORD KNOWLEDGE AND POLYSEMY IN RUSSIAN EVERYDAY LIFE LEXICON

Levin I. (levinivanse@gmail.com), **Andriyanets V.** (blindedbysunshine@gmail.com),
National Research University "Higher School of Economics";
Iomdin B. (iomdin@ruslang.ru), V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences; National Research University "Higher School of Economics";
Ambartsumian A. (anna.ambr@yandex.ru), Russian State University for the Humanities

Many words that according to the dictionaries have just one meaning are in fact understood in different ways by different speakers. In this article we deal with Russian nouns denoting everyday life objects which are subject to much variation by age, gender, and region and are poorly described by the existing dictionaries. We report the results of a multilevel survey, propose some possible metrics of word knowledge and show to what extent the words we studied are known among a certain population. We also claim that different speakers possess different sets of meanings for each word, propose ways to discover the distribution patterns for these sets and introduce the notion of disperse polysemy. We believe that our findings may be useful in lexicography (providing detailed information on current word usage in different social groups), lexical semantics (researching meaning shifts and patterns of its distribution among speakers), and language testing (more precise detection of the vocabulary sizes both in native speakers and in language learners).

CORPUS-BASED STUDY OF NON-CANONICAL USE OF RUSSIAN INTERJECTIONS

Levontina I. B. (irina.levontina@mail.ru), Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

The paper deals with the Russian interjections (oj, oh, aj, ogo, uh, etc.), namely their non-canonical use in collocations with K-words (Wh-words), mostly kak and kakoj. This type of use demonstrates a sort of syntactic recomposition — collocations oj kak, oh kakoj, etc. function as lexical units with the meaning of high degree, high quality or big quantity, although with very specific semantic shades. The paper makes use of the corpus data (the Russian National Corpus as well as the Internet data) to discover individual properties of interjections and their historical changes. Primary interjections are described against the background of interjections derived from the words of different part of speech. It turns out that in non-canonical use of primary interjections K-word can hardly be omitted, whereas derived interjections can also function the same way even without K-word. Non-canonical use of derived interjections is, with and without K-words, is very popular in contemporary Russian, especially in slang.

THE RUSSIAN 'ABY': CORPUS-DRIVEN RESEARCH (SYNCHRONY AND DIACHRONY)

Levontina I. B. (irina.levontina@mail.ru), Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia; **Shmelev A. D.** (shmelev.alexei@gmail.com)

Moscow Pedagogical State University, Moscow, Russia; Vinogradov Institute of Russian Language, Russian Academy of Sciences, Moscow, Russia; St Tikhon's Orthodox University

The paper deals with the Russian aby as a marker of "free choice" (or, rather, not specified choice criteria) within indefinite pronouns against the background of other markers of "free choice" such as ugodno, popalo, pridetsia. It pays attention not only to the synchronic semantics of aby, but also to its history and claims that the modern meaning of aby is related to its usage as a conjunction. The paper makes use of the corpus data (the Russian National Corpus as well as the Internet data) to follow the changes in the use of the particle in question over the last two hundred years. It investigates into the range of K-words that can collocate with aby: the most typical are collocations with kto, chto, kak and kakoi; however, collocations with other K-words are also present in the corpora. In addition, it discusses the question of negative polarity of aby and the increasing degree of its polarization.

AN EXPERIENCE OF THE OBJECTIVE ESTIMATION OF INTONATION QUALITY OF THE SYNTHESIZED RUSSIAN SPEECH

Lobanov B. M. (Lobanov@newman.bas-net.by), **Solomennik A. I.** (anna.i.prodan@gmail.com), **Zhitko V. A.** (zhitko.vladimir@gmail.com), United Institute of Informatics Problems NAS Belarus, Minsk, Belarus

The paper describes an experiment on an instrumental evaluation of the intonation quality of synthesized Russian speech by using of "Inton@Trainer" computer system. The system was originally designed to train learners in producing the basic intonation patterns of Russian speech. It is based on comparing the melodic portraits of a reference sentence and a sentence pronounced by the learner. Our approach to assessing the intonational quality of speech allows to treat a synthesized speech with the same strict requirements as are applied to students studying Russian as a second language. We describe the technology used for the instrumental evaluation of the intonation quality of synthesized speech and the acoustic database of reference phrases used to assess the intonation quality of synthesized speech. The paper presents the results of testing the intonation quality of two Russian synthetic voices. We discuss the results of the experiment and outline the ways for improving the methods for objective evaluation of synthesized speech prosodic quality, as well as the possibility of applying the developed system in other linguistic tasks.

EXTRACTING SENTIMENT ATTITUDES FROM ANALYTICAL TEXTS

Loukachevitch N. V. (louk_nat@mail.ru), Lomonosov Moscow State University, Moscow, Russia; **Rusnachenko N.** (kolyarus@yandex.ru), Bauman Moscow State Technical University, Moscow, Russia

In this paper we present the RuSentRel corpus including analytical texts in the sphere of international relations. For each document we annotated sentiments from the author to mentioned named entities, and sentiments of relations between mentioned entities. In the current experiments, we considered the problem of extracting sentiment relations between entities for the whole documents as a three-class machine learning task. We experimented with conventional machine-learning methods (Naive Bayes, SVM, Random Forest).

RE-INTERPRETING EVENTS: NOTES ON ONE LINGUISTIC INNOVATION IN RUSSIAN

Lyutikova E. A. (lyutikova2008@gmail.com), **Tatevosov S. G.** (tatevosov@gmail.com), Lomonosov Moscow State University, Moscow Pedagogical State University, Moscow, Russia

The paper explores the distribution and interpretation of the discourse marker po(-)xodu (PX) and addresses a possible path of its diachronic development. We argue that the range of uses of PX attested in the corpora supports an analysis that identifies three meanings / functions of

this item labeled eventive PX, epistemic PX and discourse-level PX throughout this paper. We propose that the latter two are the products of re-interpretation of the former. We argue for a presuppositional analysis of the eventive PX whereby it requires there be a set of background events that show a temporal overlap with the asserted event and add up to the integral whole. We analyze the epistemic PX as resulting from inferential reinterpretation of the relationship between background and asserted events, with the abductive reasoning being the key ingredient of this reinterpretation. Finally, we treat the discourse-level PX as a counterpart of the eventive PX in the domain of speech acts. We speculate that Krifka's (2014) recent view of speech acts as index changers opens a way of accounting for this parallelism in a principled way. On the diachronic side, we identify PX as the product of diachronic development of the construction in which the argument of the noun xod 'move' is expressed by an overt DP. In the course of development, this DP was first replaced by pro, which gave rise to the eventive PX, and later on developed epistemic and discourse-level meanings / functions.

LEVERAGING DEEP NEURAL NETWORKS AND SEMANTIC SIMILARITY MEASURES FOR MEDICAL CONCEPT NORMALISATION IN USER REVIEWS

Miftahutdinov Z. (zulfatmi@gmail.com), **Tutubalina E.** (elvtutubalina@kpfu.ru),
Kazan Federal University, Kazan, Russia

Nowadays a new yet powerful tool for drug repurposing and hypothesis generation emerged. Text mining of different domains like scientific libraries or social media has proven to be reliable in that application. One particular task in that area is medical concept normalization, i.e. mapping a disease mention to a concept in a controlled vocabulary, like Unified Medical Language System (UMLS). This task is challenging due to the differences in language of health care professionals and social media users. To bridge this gap, we developed end-to-end architectures based on bidirectional Long Short-Term Memory and Gated Recurrent Units. In addition, we combined an attention mechanism with our model. We have done an exploratory study on hyperparameters of proposed architectures and compared them with the effective baseline for classification based on convolutional neural networks. A qualitative examination of the mentions in user reviews dataset collected from popular online health information platforms as well as quantitative one both show improvements in the semantic representation of health-related expressions in user reviews about drugs.

MACHINE LEARNING CLASSIFICATION OF USER INTERESTS ACROSS LANGUAGES AND SOCIAL NETWORKS

Mikhalkova E. V. (e.v.mikhalkova@utmn.ru), **Ganzherli N. V.** (n.v.ganzherli@utmn.ru),
Karyakin Y. E. (y.e.karyakin@utmn.ru), **Grigoryev D. A.** (Grigd2013@gmail.com),
Tyumen State University (University of Tyumen), Tyumen, Russia

Being a matter of cognition, user interests should be apt to classification independent of the language of users, social network and the essence of interest itself. To prove it, we built a collection of English and Russian Twitter and Vkontakte community pages manually classified according to the interests of their followers. First, we created a model of Major Interests (MaIs) with the help of expert analysis and then classified the mentioned set of pages using machine learning algorithms (SVM, Neural Network, Naive Bayes, Logistic Regression, Decision Trees, k-Nearest Neighbors) trying different optimization techniques. We take three interest domains that are typical of both English and Russian-speaking communities: football, rock music, vegetarianism. The results of classification show a greater correlation between Russian-Twitter and English-Twitter pages. The Logistic Regression with Bernoulli bag-of-words model proves to be the most effective classification algorithm.

ANALYSIS OF COREFERENTIAL EXPRESSIONS IN PAWS (ENGLISH-CZECH-RUSSIAN-POLISH PARALLEL TREEBANK WITH ANAPHORIC RELATIONS)

Nedoluzhko A. (nedoluzko@ufal.mff.cuni.cz)[1], **Novák M.** (mnovak@ufal.mff.cuni.cz)[1],
Ogrodniczuk M. (maciej.ogrodniczuk@ipipan.waw.pl)[2];
[1]Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic; [2]Polish Academy of Sciences, Institute of Computer Science, Warsaw, Poland

In this paper, we decribe the coreference annotation on a multi-lingual parallel treebank (PAWS), a portion of Wall Street Journal translated into Czech, Russian and Polish which continues the tradition of multilingual treebanks with coreference annotation. The paper focuses on language-specific differences. We analyse syntactic structures concerning anaphoric relations in the languages under analysis, such as personal and impersonal constructions in poly-predicative constructions and pro-drop qualities.

PRONOMINAL ADVERBS IN GERMAN AND THEIR EQUIVALENTS IN ENGLISH, CZECH AND RUSSIAN: EVIDENCE FROM THE PARALLEL CORPUS

Nedoluzhko A. (nedoluzko@ufal.mff.cuni.cz), Charles University, Czech Republic;
Lapshinova-Koltunski E. (e.lapshinova@mx.uni-saarland.de), Saarland University, Germany

The paper presents a contrastive analysis of pronominal adverbs in German (dabei, darauf, damit etc.) and their equivalents in English, Czech and Russian. The analysis is based on an empirical study of parallel news texts. Our main focus is to show the interplay between cohesive devices expressed through German pronominal adverbs in text and explore their equivalents in English, Czech and Russian. As the dataset at hand contains translations, we also focus on the influence of the translation factor in parallel texts.

SUSPENDED ASSERTION AND NONVERIDICALITY

Paducheva E. V. (elena.paducheva@yandex.ru),
Informatics and Control Federal Research Centre RAN, Moscow, Russia

The paper addresses the notion of “snyataya utverditel’nost’” (suspended assertion). The author argues that the term “suspended assertion”, introduced by U. Weinreich in 1963, covers the same range of phenomena as the term nonveridicality (its suggested Russian equivalent is neveridicativnost’), which has become widespread due to the works by F. Zwarz, A. Giannakidou and many others. It is demonstrated that the notion of suspended assertion an be applied to interpret a number of facts of the Russian language, such as nibud’-pronouns, pronouns of negative polarity, the disappearance of a semantic argument of verbs with the direct (non- parametrical) diathesis, the mirror symmetry of past and future, the negation with an extended scope, nibud’-pronouns in the scope of negation, the interchangeability of eshche ‘yet’ and uzhe ‘already’. It’s the author’s conviction that the notion of suspended assertion will be applicable in many other contexts.

RUSSE'2018: A SHARED TASK ON WORD SENSE INDUCTION FOR THE RUSSIAN LANGUAGE

Panchenko A. (panchenko@informatik.uni-hamburg.de), University of Hamburg, Hamburg, Germany; **Lopukhina A.** (alopukhina@hse.ru), National Research University Higher School of Economics, Moscow, Russia; The Vinogradov Institute of the Russian Language, Russian Academy of Sciences, Moscow, Russia; **Ustalov D.** (dmitry@informatik.uni-mannheim.de), University of Mannheim, Mannheim, Germany; Krasovskii Institute of Mathematics and Mechanics, Yekaterinburg, Russia; **Lopukhin K.** (kostia.lopuhin@gmail.com), Scrapinghub, Moscow, Russia; **Arefyev N.** (nick.arefyev@gmail.com), Lomonosov Moscow State University, Moscow, Russia; Samsung Moscow Research Center, Moscow, Russia; **Leontyev A.** (aleksey_l@abbyy.com), ABBYY, Moscow, Russia; **Loukachevitch N.** (louk_nat@mail.ru), Lomonosov Moscow State University, Moscow, Russia

The paper describes the results of the first shared task on word sense induction (WSI) for the Russian language. While similar shared tasks were conducted in the past for some Romance and Germanic languages, we explore the performance of sense induction and disambiguation methods for a Slavic language that shares many features with other Slavic languages, such as rich morphology and virtually free word order. The participants were asked to group contexts of a given word in accordance with its senses which were not provided beforehand. For instance, given a word "bank" and a set of contexts for this word, e.g. "*bank* is a financial institution that accepts deposits" and "river *bank* is a slope beside a body of water", a participant was asked to cluster such contexts into *unknown in advance* number of clusters corresponding to, in this case, the "company" and the "area" senses of the word "bank". For the purpose of this evaluation campaign, we developed three new evaluation datasets based on sense inventories that have different sense granularity. The contexts in these datasets were sampled from texts of Wikipedia, the academic corpus of Russian, and an explanatory dictionary of Russian. Overall, 18 teams participated in the competition submitting 383 models. Multiple teams managed to substantially outperform competitive state-of-the-art baselines from the previous years based on sense embeddings.

SPEECH ACT CONJUNCTION: THE SCALE OF SPEECH ACT USE AND ITS MANIFESTATION IN GRAMMAR

Pekelis O. E. (opekelis@gmail.com), Russian State University for the Humanities, Moscow, Russia

This paper deals with the phenomenon of speech act conjunction in which the relation expressed by the conjunction holds on the level of speech act performance rather than on the level of states of affairs. It is argued that besides clearly speech act and clearly non-speech act uses, there is a class of constructions of an intermediate nature. The criteria are proposed that serve to distinguish between these three types of use. In particular, it is demonstrated that imperative sentences can only be of the "intermediate" type, while interrogative sentences can represent the clearly speech act use. The proposed distinction manifests itself in grammar. Namely, different conjunctions are compatible with different types of speech act use; the correlative item togda ('then') cannot be used within a clearly speech act construction.

SEMI-AUTOMATIC INTEGRATION OF A NEW LANGUAGE INTO A MULTILINGUAL NLP MODEL: THE CASE OF JAPANESE

Petrova M. A. (maria_pet@abbyy.com), **Druzhkina A. A.** (anna_r@abbyy.com), **Garashchuk R. V.** (ruslan_g@abbyy.com), **Yudina M. V.** (maria_yu@abbyy.com), ABBYY, Moscow, Russia

The current paper deals with the integration of the Japanese language in a multilingual NLP model, namely, the Compreno model. The formalism includes morphological, syntactic and semantic patterns, covering all possible semantic and syntactic dependencies a word can attach. The architecture of the model allows us to acquire nearly all semantic links of a word through its proper positioning in a thesaurus-like semantic hierarchy, where words are linked through semantic dependencies. The inheritance principle of the hierarchy simplifies the syntactic description of a newly added language as well. Unlike the traditional approach to Japanese parsing

based on chunks, or bunsetsus, we suggest a Japanese parser based on constituents. Special attention is given to the tools that allow us to automatize language description process and significantly speed up the description. The work on the Japanese model is still in progress, therefore, we show the current results we have achieved, and point out problems that remain to be solved.

CORPUS SIZE AND THE ROBUSTNESS OF MEASURES OF CORPUS DISTANCE

Piperski A. Ch. (apiperski@gmail.com), Russian State University for the Humanities / National Research University Higher School of Economics, Moscow, Russia

This paper studies the impact corpus size has on the robustness of various frequency-based measures of corpus distance (or similarity, respectively), such as Euclidean distance, Manhattan distance, Cosine distance, χ^2, Spearman's ρ, and Simple-Maths Keyword distance. An experiment performed using the British National Corpus shows that Euclidean distance is least influenced by corpus size and thus is best suited for the purpose of comparing corpora.

"A U NAS V KVARTIRE GAZ. A U VAS?": THE RUSSIAN CONJUNCTION A VIEWED THROUGH THE PRISM OF PROSODICALLY ANNOTATED CORPUS DATA

Podlesskaya V. I. (vi_podlesskaya@il-rggu.ru), Russian State University for the Humanities, Russian Academy of National Economy and Public Administration; Moscow, Russia

The paper focuses on Russian constructions with clauses (or VPs) combined by means of the discourse marker A, that behaves as a conjunction or as a particle in different contexts. Prosodically, the construction may come up in two forms: (a) as a single illocution with the first clause pronounced with a rising pitch that projects discourse continuation, and (b) as two separate illocutions with the first clause pronounced with a falling pitch that projects no continuation. Basing on the data from the Prosodically Annotated Corpus of Spoken Russian, prosody and grammar of (a) and (b) were analyzed qualitatively and quantitatively. Type (b) appeared to be as frequent as type (a) and systematically favored in pragmatically marked contexts.

REFERRING EXPRESSION GENERATION FOR QUESTION ANSWERING AND GRAPH VISUALIZATION

Rygaev I. P. (irygaev@gmail.com), Laboratory of Computational Linguistics, A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia

This paper describes a practical solution for the task of referring expressions generation (REG) in the context of a question-answering system. When an answer to a question is found in the knowledge base the system has to decide how to present the answer to the user, which properties uniquely distinguish the object found from other objects in the knowledge base. Another task where referring expressions would be useful is the semantic graph visualization task. Building on top of the graph-based approach presented by Krahmer et al in 2003 this paper provides some practical improvements to the algorithm, namely: 1) Instead of depth-first graph search we use breadth-first search, which is dramatically faster when a scene graph is big but the description graph to be found is small, 2) Limit on the size (the number of edges) of the resulting description graph to increase performance and avoid useless long descriptions. Also a sketch on linguistic realization of the referring expressions is outlined.

THE STRUCTURE OF EVERYDAY DIALOGUE AS THE SEQUENCE OF SPEECH ACTS

Sherstinova T. Yu. (t.sherstinova@spbu.ru), Philological Faculty of Saint-Petersburg State University; National Research University Higher School of Economics; Saint-Petersburg, Russia

The structure of Russian everyday dialogue was studied on the basis of 73 microdialogues of everyday speech communication from the "One Day of Speech" corpus (the ORD Corpus). The aim of the research was to find out what types of speech acts commonly initiate and complete

everyday dialogues, as well as to reveal the most typical sequences of speech acts in these dialogues. Altogether, 2230 speech acts of 30 people referring to both professional, and household conversations have been analysed. N-gram analysis has been used to calculate the most frequent sequences of speech acts. The obtained results showed that dialogues are usually started by representatives, i. e. speech acts related to the exchange of information (38% of all cases), etiquette beginnings (greetings, vocatives) take place in 23% of the dialogues, and in 19% of cases the conversation begins with a regulative form. Speech acts ending dialogues show a greater variety: representatives contribute 2% of all dialogue ends, valuative judgments and regulatory forms cover 14% each, further go directives (8%), commissions (8%), etiquette forms (8%) and emotional and expressive form (7%). As for the most typical bigrams of speech acts, they are the following: two consecutive representatives (22.35%), a regulatory form followed by a representative (6.93%), a representative and a regulatory form (6%), a valuative with a following representative (5.21%), a representative and a valuative judgment (4.77%), as well as two combinations of a directive with a representative (2.77% each). Besides, the article presents data on the occurrence of the most frequent pairs of speech acts at the subtype level. Here, the most frequent one is the sequence "question'+"answer', which covers 2.45%.

IMPROVING TOPIC MODELS WITH SEGMENTAL STRUCTURE OF TEXTS

Skachkov N. A. (nikolaj-skachkov@yandex.ru), Lomonosov Moscow State University;
Vorontsov K. V. (voron@forecsys.ru), Dorodnicyn Computing Centre of RAS, Moscow Institute of Physics and Technology, Russia

Probabilistic topic modeling is a powerful tool of text analysis, that reveals topics as distributions over words and then softly assigns documents to the topics. Even though the aggregated distributions can be good with basic models, a sequential topic representation of each document is often unsatisfactory. This work introduces a method that allows to increase the quality of topical representation of each single text using its segmental structure. Our approach is based on Additive Regularization of Topic Models (ARTM), which is a technique for imposing additional criteria into the model. The proposed method efficiently avoids a bag-of-words assumption by considering the topical connections of words that co-occur in a local segment. We assume, that sequential sentences are topically and semantically coherent, while the number of topics in each particular text fragment is low. We apply our model to topic segmentation task and achieve a better quality than the current state-of-the-art TopicTiling algorithm. In further experiments we demonstrate that the proposed technique reveals an interpretable sequential structure of documents, while keeping a number of topics low, i.e. the sparsity of the model increases. Apart from topic segmentation, the constructed topical text embeddings can be used in any other applications, where the analysis of the document structure is desirable.

BUILDING A CORPUS FOR THE QUANTITATIVE RESEARCH OF RUSSIAN DRAMA: COMPOSITION, STRUCTURE, CASE STUDIES

Skorinkin D. (dskorinkin@hse.ru), **Fischer F.** (frafis@gmail.com),
Palchikov G. (rebel368@gmail.com), National Research University Higher School of Economics

In this paper we introduce RusDraCor—an open corpus of Russian drama for digital literary & linguistic research. The corpus (rus.dracor.org) contains plays from the middle of XVIII to the first third of XX century provided with structural (plus some semantic) markup and metadata. Texts are encoded in the XML-based standard TEI, widely used in building corpora for the humanities. We describe the contents and annotation layers of our corpus, provide some details on its development and enrichment, and finally describe three research cases. Each case demonstrates the use of RusDraCor to answer specific questions about composition, structural features and historical evolution of Russian drama.

SPEECH DISFLUENCIES ANALYSIS IN THE DISCOURSE OF 10–12 YEARS OLD NATIVE RUSSIAN SPEAKING CHILDREN

Slabodkina T. A. (slabodkina.t@gmail.com), Lomonosov Moscow State University, Moscow, Russia; **Fedorova O. V.** (olga.fedorova@msu.ru), Lomonosov Moscow State University, Institute of Linguistics RAS, RANEPA, Moscow, Russia

The paper reviews the problem of speech disfluency which over the years has become traditional for the "Dialogue" conference (see Podlesskaya, Komarova 2010; Laurinavichyute, Fedorova 2010; Fedorova 2010; Podlesskaya 2013; Bogdanova-Beglarian 2013; Podlesskaya 2014; Potanina et al. 2016). In this paper, we compared speech disfluencies in two corpora of dialogues between children of 10–12 years old (section 1) and adults (section 2). Both corpora were collected using the referential communication task "Tangrams" (to perform the task, participants had to agree on the nomination of some abstract figures).

In the third section of the text, the authors provide the classifications of speech disfluencies present in the dialogues with examples. The results of the comparison and the methods of analysis are given in the fourth paragraph. Finally, the last section contains the discussion of the results and perspectives of the further work. The paper shows that speech of children of the given age group differs from adults' speech in terms of disfluencies at the discourse level.

GENDER, DECLENSION AND STEM-FINAL CONSONANTS: AN EXPERIMENTAL STUDY OF GENDER AGREEMENT IN RUSSIAN

Slioussar N. A. (slioussar@gmail.com), Higher School of Economics, Moscow, and Saint-Petersburg State University, Saint-Petersburg, Russia

Every adult native speaker of Russian knows that kon' is masculine and lan' is feminine, although 3rd declension nouns present some difficulties in the first and second language acquisition. However, will the fact that these nouns are less frequent than masculine nouns ending in a consonant or feminine nouns ending in -a/ja play a role for online subject-predicate agreement processing? Or will subject-predicate agreement processing be more problematic with subjects of a certain gender? Finally, some final consonants are more characteristic for feminine gender, while the others for masculine gender. Are speakers sensitive to this? We present two experiments addressing these questions. We found that all three factors play a role, but for different tasks (online agreement processing or determining the gender of a novel word) and at different processing stages.

IMPROVING NEURAL MORPHOLOGICAL TAGGING USING LANGUAGE MODELS

Sorokin A. A. (alexey.sorokin@list.ru), Moscow Institute of Physics and Technology, Dolgoprudnyj, Russia; Lomonosov Moscow State University, Moscow, Russia

We offer a new neural architecture for character-level morphological tagging, combining character-level networks with the output of neural language model on morhological tags. Our proposal reduces tagging error up to 10% in comparison with baseline model and achieves state-of-the-art performance both on ru_syntagrus and MorphoRuEval datasets.

DIFFERENTIAL OBJECT MARKING IN CONTACT-INFLUENCED RUSSIAN SPEECH: EVIDENCE FROM THE CORPUS OF CONTACT-INFLUENCED RUSSIAN SPEECH OF RUSSIAN FAR EAST AND NORTHERN SIBERIA

Stoynova N. M. (stoynova@yandex.ru), Vinogradov Russian Language Institute, RAS; Moscow, Russia

The paper deals with differential object marking in the Russian Speech of Nanai-Russian bilingual speakers, namely the variation such as принес рыбу ~ принес рыба ('{he} brought fish-acc ~ fish-nom'). The puzzle is that this peculiarity can result from a number of different processes: morphosyntactic borrowing from Nanai, penetration of dialectal features into the speech of bilinguals, under-acquisition or reinterpretation of the Standard Russian system. The data of a small corpus of contact-influenced Russian Speech is used to test all these hypotheses. The results are following. Nominative forms are used in DO-position in quite a systematic way and such uses cannot be

estimated as occasional "errors". The main factors that influence the NOM~ACC distribution are a) information structure and b) the accentual type of noun stem. The latter fact supports the hypothesis of a systematic reinterpetation of the Standard Russian system in the situation of incomplete acquisition. No significant correlations with animacy, definiteness, verb form and word order were attested. DOM pattern of Nanai Russian differs from those of Russian dialects and reveals some similarity to those of Nanai. However it cannot be considered as a full morphosyntactic calque.

THE INTERPRETATION OF RUSSIAN PRONOUNS IN COUNTERIDENTITY CONTEXTS: A CORPUS STUDY

Tiskin D. B. (daniel.tiskin@gmail.com), Saint Petersburg State University

This paper is a first step towards a corpus-based description of the semantics of Russian pronouns in intensional contexts. Having justified the use of corpus in (formal) semantic research, I delineate a particular issue within the topic: whether a given pronoun is interpreted de se or de re in counteridentity contexts. A counteridentity context is a clause within the scope of a counterfactual (clause or adverbial) that affects the identity of a real individual, e.g. if I were you, were I you, etc. If a pronoun such as I, my or the Russian reflexive possessive svoj is used in such a context, two options are theoretically possible: either it picks out the speaker's real self (de re), or it refers to the identity assumed by the speaker in the contrary-to-fact situations introduced by the counterfactual (de se). Using data from the GICR corpus (approx. 20 billion tokens), I show that for the Russian first-person singular pronoun ja and its corresponding possessive moj, de se reference is possible but de re interpretation is more frequent. The opposite holds for the reflexive sebja, whereas svoj is interpreted de se with no exception. Special attention is paid to situations where more than one referential strategy is possible. The paper concludes with a couple of observations relevant for the future formal accounts of de se reference.

THE CUES FOR RHETORICAL RELATIONS IN RUSSIAN: "CAUSE—EFFECT" RELATION IN RUSSIAN RHETORICAL STRUCTURE TREEBANK

Toldova S. (toldova@yandex.ru), NRU Higher School of Economics, Moscow, Russia; **Pisarevskaya D.** (dinabpr@gmail.com), **Kobozeva M.** (kobozeva@isa.ru); Institute for Systems Analysis FRC CSC RAS, Moscow, Russia; **Vasilyeva M.** (linellea@yandex.ru), Lomonosov Moscow State University, Moscow, Russia

The purpose of the paper is to investigate cues signalling the relations between discourse units in Russian. Building a lexicon of discourse connectives is an indispensable subtask in many discourse parsing applications as well as an essential issue in theoretical researches of text coherence. In order to develop such a resource for Russian, we have conducted a corpus-based study of discourse connectives that were manually extracted from the Russian Rhetorical Structure Treebank (Ru-RSTreebank). The Treebank includes 79 texts annotated within the RST framework [Mann, Thompson 1988]. In order to provide a deeper analysis of connectives in Russian, we focus on causal relations only, namely, the 'Cause-Effect' relation. Some of the connectives (primary connectives) are enumerated in grammars and dictionaries. They primarily mark the intra-sentential relations. However, there is an expansive class of less grammaticalized items (secondary connectives) that have received less attention till now. Some of them are based on content words (e.g. по причине 'for the cause'). Secondary connectives often serve as linking devices for inter-sentential relations. We suggest a scheme for connectives annotation for Russian. We specify the basic patterns that can be used for less-grammaticalized connectives mining in an unannotated corpus. Besides, we provide the comparison of two classes of connectives (primary vs. secondary ones). Our research has shown that these two classes differ in their properties. There is a statistically significant difference between them with respect to the nucleus/satellite position, intra- vs. inter-sentential relations and some others.

SYNTAX OF PREPOSITIONAL ADVERBS: SOME DIFFICULT CASES

Uryson E. V. (uryson@gmail.com), Russian Language Institute (RAS), Moscow, Russia

The subject of this paper are Russian so called adverbial prepositions; cf. vokrug (kostra) 'around smth.', daleko ot (doma) 'far from smth.', etc. By definition, an adverbial preposition either coincides with an adverb (cf. vokrug) or contains an adverb and a preposition (cf. daleko ot). As I have

demonstrated in my previous works, an adverbial preposition and the underlying adverb have the same meaning, the only difference between them being in 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'. From the modern point of view, syntactic distinction is insufficient for interpreting such cases as different words (or different meanings of a word). So, an adverbial preposition and the underlying adverb should be interpreted as the same meaning of a given word. I argue that this word is an adverb (or a prepositional adverb). This paper deals with syntax of these adverbs. Such adverbs have one or more semantic actants, at least one of them being expressed by a noun or a prepositional group. The problem is that in some cases it is not clear whether the prepositional group is governed by the adverb or by the verb governing this adverb (thus the adverb and the prepositional group are co-governed by the verb). A criterion of adverb vs. verb governing of such groups is discussed. Two Russian adverbs zadolgo 'for a long time before smth.' and nezadolgo 'for a long time before smth.' are described from this point of view.

'CHTO BUDET, TO (I) BUDET': ON ONE PATTERN OF TAUTOLOGIES IN RUSSIAN

Vilinbakhova E. L. (e.vilinbakhova@spbu.ru), St. Petersburg State University, St. Petersburg, Russia

This paper contributes to the debate on the analysis of linguistic tautologies—structures that state an unquestionable truth by virtue of their logical form and therefore require a reinterpretation to be informative. While there is a great number of studies of nominal tautologies of the form 'X is X', clausal tautologies, i. e. conditionals 'if P, P', disjunctives 'either P or not P', free relatives 'P, what P', etc., are given less attention. This paper investigates one of such patterns, namely, correlative tautologies, where the subordinate clause precedes the main clause, that could be exemplified by the expression chto budet to (i) budet lit. 'what will be that (EMPH) will be'. The data taken from the Russian National Corpus and Internet as well as dictionary definitions show that tautologies of this kind exhibit various peculiar properties. First, some correlative tautologies can receive opposite interpretations in different contexts, i. e. chto bylo, to bylo lit.'what has been that has been' can mean both 'this fact cannot be denied' [Bylugina, Shmelev 1997] or 'the past should be forgotten for the sake of the future' [Active Dictionary of Russian]. Next, the particle i, which is commonly used in Russian correlatives, cf. [Mitrenina 2010], is acceptable for some tautologies but not licensed in others. I argue that for correlative tautologies the crucial ingredient is salience of the situation in question as presented by the speaker that, along with specific vs. generic readings available, results in four possible strategies of their interpretation.

IMPERATIVES, VOCATIVES, AND QUESTIONS IN COHERENT DISCOURSE: THE PROSODIC MARKERS OF INCOMPLETENESS IN THE RUSSIAN SPOKEN SPEECH CORPORA

Yanko T. E. (tanya_yanko@list.ru), Institute of Linguistics, Moscow, Russia

One of the means of designating the coherence in the spoken discourse is demonstrating that the current utterance of the discourse is not terminal. Every step of narrative consisting of the chain of statements can be marked as non-final. The prosodic cues for incompleteness applied to the speech act of a statement have been studied in details in linguistic literature. In this paper, the discourse incompleteness is analyzed as composed not only with statements but with questions, imperatives, and vocatives as well. The results of the investigation are as follows. The wh-questions, imperatives, and vocatives can be freely composed with the meaning of discourse continuity, and they have specific prosodic cues for marking this combination of meanings. Whereas the yes-no-questions do not accept the prosodic incompleteness marking. The prosodic patterns of incompleteness and the accent placement in questions, vocatives, and imperatives are exemplified here by the dialogues taken from the Multimodal corpus of the Russian National corpus, the Prosodically Annotated Corpus of Spoken Russian (spokencorpora.ru), and the minor working collection of the Russian speech recordings specifically set up for this investigation. The software program Praat was used in the process of analyzing the sounding data.

RUSSIAN 'KAK-NIBUD'' THROUGH THE PRISM OF PARALLEL CORPORA

Zalizniak Anna. A. (anna.zalizniak@gmail.com), Institute of Linguistics of the RAS; Institute of Informatics Problems of the FRC CSC RAS, Moscow, Russia;
Denisova G. V. (galina.denissova@unipi.it), Pisa State University, Italy;
Mikaelian I. L. (irina-mikaelian@yandex.ru), Pennsylvania State University, USA

The paper proposes a semantic analysis of the Russian indefinite adverb kak-nibud' based on the data collected from the French-Russian, Italian-Russian, and English-Russian parallel subcorpora of the Russian National Corpus, as well as from the Data Base of the Russian Discourse Markers and their French equivalents. The study applies the "unidirectional method" of contrastive analysis within which the translation by a professional translator is viewed as a quasi-lexicographic explication of a given unit revealing implicit components of its semantics. Our analysis demonstrates that kak-nibud' is a highly language-specific Russian word. It reflects in a high percentage of null equivalents of this unit in the three languages under investigation, for both Russian taken as the source or target language. The study has also allowed us to show that the analyzed adverb can function as a marker of non-controllability of a hypothetic event similar to the function of the subjunctive mood in Romance languages. On the other hand, the use of kak-nibud' ('anyhow', 'poorly') in a purely evaluative meaning cited by monolingual and bilingual dictionaries has shrunk in contemporary Russian compared to the Russian of the 19th century.

TWO DIALECTS OF RUSSIAN GRAMMAR: CORPUS DATA AND FORMAL MODELS

Zimmerling A. V. (fagraey64@hotmail.com), Pushkin State Russian Language Institute, Moscow State University of Education, Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia

This paper is addressed the problem of parametric variation in Russian grammar, with focus on copular constructions with agreeing and non-agreeing adjectival predicates. Basing on Russian National Corpus, I reconstruct two dialects of Russian morphosyntax. They differ regarding the assignment of the predicative instrumental case, raising conditions and the distribution of agreeing vs non-agreeing predicates after быть 'be', стать 'become' and казаться 'seem'. Russian-A only licenses predicative instrumental on adjectives after SEEM (казалось странным, что P) and non-agreeing predicatives after non-zero forms of BE or BECOME (было странно, что P). Russian-B allows non-agreeing forms after SEEM (казалось странно, что P) and forms of the predicative instrumental case after non-zero forms of BE and BECOME (было странным, что P). I argue that the differences between Russian-A and Russian-B must explained in terms of parametric settings and claim that Russian predicatives lack forms of the predicative instrumental. The assignment of the predicative instrumental to adjectival heads can be explained as subject control in all dialects, but only Russian-B allows raising of sententional arguments to the position of the matrix subject.

DEVELOPMENT OF COMMUNICATIVE BEHAVIOR MODEL FOR F-2 ROBOT BASING ON "REC" MULTIMODAL CORPORA

Zinina A. A. (zinina_aa@nrcki.ru), **Arinkin N. A.** (arinkin_na@nrcki.ru), **Zaydelman L. Ya.** (zaydelman_ly@nrcki.ru), **Kotov A. A.** (kotov_aa@nrcki.ru), Kurhcatov Institute, Moscow, Russia

The article describes the developed architecture for modeling natural communicative behavior on the F-2 robot. The important part of our work is the study of human communicative behavior and the transfer of this behavior to the robot. For this purpose we are developing the Russian Emotional Corpus (REC) where video recordings of natural emotional dialogues are collected. We explore the features of natural communication, and also develop an architecture that takes into account these features. For example, using the architecture presented in the article a robot can express any communicative function, using one or more executive organs: for example, to express an appeal with facial expressions, head movements or gestures. The developed architecture also allows us to flexibly combine gestures with different communicative functions. The architecture allows us to use "split", "join" and "single" modes to combine tags from different BML-packages, and also to synchronize tags in a single BML-package. These features are important for modeling of human-like behavior for the robot F-2, and are necessary to improve the communication between a robot and a user.

Авторский указатель

Author Index

Научное издание

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной
международной конференции «Диалог»

Выпуск 17 (24). 2018

Ответственный за выпуск **А. В. Ульянова**
Вёрстка **К. А. Климентовский**

Издательский центр «Российский
государственный гуманитарный университет»
125993, Москва, Миусская пл., д. 6
Тел.: +7 499 973 42 06